INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/13295

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C07H 21/04
US CL : 435/6; 536/24.33
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 172.3; 536/24.33

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,180,673 A (WILSON et al) 19 January 1993, see entire document.	1-17
Y	US 5,492,823 A (XU) 20 February 1996, see entire document.	1-17
Y	US 5,472,872 A (MEAD et al) 05 December 1995, see entire document.	1-17
Y	US 5,491,060 A (STEIN et al) 13 February 1996, see entire document.	1-17
Y	US 5,563,328 A (MITRA et al) 08 October 1996, see entire document.	1-17
Y	US 5,677,180 A (ROBINSON et al) 14 October 1997, see entire document.	1-17

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 AUGUST 1999	19 OCT 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
ARUN CHAKRABARTI
Telephone No. (703) 306-5818

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/13295

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 5,830,731 A (SEED et al) 03 November 1998, see entire document.	1-17
Y,P	US 5,770,420 A (LOWE et al) 23 June 1998, see entire document.	1-17
Y	PIEKAROWICZ. A.et al, A New Method for the Rapid Identification of Genes Encoding Restriction and Modification Enzymes. Nucleic Acids Research. March 1991. Vol. 19. No. 8., pages 1831-1835.	1-17

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, MEDLINE, BIOSIS, CAPLUS, EMBASE, GENBANK
search terms: method, cloning, genes, cell surface antigen, transporter, gene cassettes, restriction endonuclease, methyltransferase, enzymes

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04

A1

(11) International Publication Number: WO 99/64632

(43) International Publication Date: 16 December 1999 (16.12.99)

(21) International Application Number: PCT/US99/13295

(22) International Filing Date: 11 June 1999 (11.06.99)

(30) Priority Data:

60/089,086	12 June 1998 (12.06.98)	US
60/089,101	12 June 1998 (12.06.98)	US

(71) Applicant *(for all designated States except US)*: NEW ENGLAND BIOLABS, INC. [US/US]; 32 Tozer Road, Beverly, MA 01915 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: RALEIGH, Elisabeth, A. [US/US]; 32 Barton Street, Somerville, MA 02144 (US). VAISVILA, Romualdas [LT/LT]; 507 Sandy Bay Terrace, Rockport, MA 01966 (US). MORGAN, Richard, D. [US/US]; 31 Donovan's Way, Middleton, MA 01949 (US).

(74) Agent: WILLIAMS, Gregory, D.; New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: RESTRICTION ENZYME GENE DISCOVERY METHOD

(57) Abstract

The invention is directed to direct cloning of intact genes, with a high probability that the orientation of expression is known in advance, and with a low probability of being associated with extraneous possibly toxic genes. The invention is particularly directed to obtaining genes encoded in DNA cassettes comprised of repeat sequences flanking variable open reading frames. The invention encompasses obtaining such cassette–encoded genes using oligonucleotides hybridizing to the repeated elements, cloning them and expressing them. Expression may employ tightly regulated vectors and useful strains disclosed. Methods for identifying restriction endonuclease and DNA methyltransferase genes in the absence of prior information about the sequences or biochemical specificities of these are also disclosed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Larvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/13295

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C07H 21/04
US CL : 435/6; 536/24.33
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 172.3; 536/24.33

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,180,673 A (WILSON et al) 19 January 1993, see entire document.	1-17
Y	US 5,492,823 A (XU) 20 February 1996, see entire document.	1-17
Y	US 5,472,872 A (MEAD et al) 05 December 1995, see entire document.	1-17
Y	US 5,491,060 A (STEIN et al) 13 February 1996, see entire document.	1-17
Y	US 5,563,328 A (MITRA et al) 08 October 1996, see entire document.	1-17
Y	US 5,677,180 A (ROBINSON et al) 14 October 1997, see entire document.	1-17

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

13 AUGUST 1999

Date of mailing of the international search report

19 OCT 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

ARUN CHAKRABARTI

Telephone No. (703) 306-5818

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 5,830,731 A (SEED et al) 03 November 1998, see entire document.	1-17
Y,P	US 5,770,420 A (LOWE et al) 23 June 1998, see entire document.	1-17
Y	PIEKAROWICZ. A.et al, A New Method for the Rapid Identification of Genes Encoding Restriction and Modification Enzymes. Nucleic Acids Research. March 1991. Vol. 19. No. 8., pages 1831-1835.	1-17

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, MEDLINE, BIOSIS, CAPLUS, EMBASE, GENBANK
search terms: method, cloning, genes, cell surface antigen, transporter, gene cassettes, restriction endonuclease, methyltransferase, enzymes

3

PATENT COOPERATION TREATY

PCT

REC'D 14 AUG 2000

WIPO PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference NEB-165-PCT	**FOR FURTHER ACTION**	See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/US99/13295	International filing date *(day/month/year)* 11 JUNE 1999	Priority date *(day/month/year)* 12 JUNE 1998
International Patent Classification (IPC) or national classification and IPC IPC(7): C12Q 1/68; C07H 21/04 and US Cl.: 435/6; 536/24.33		
Applicant NEW ENGLAND BIOLABS, INC.		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 4 sheets.

☐ This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority. (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

These annexes consist of a total of 0 sheets.

3. This report contains indications relating to the following items:

- I ☒ Basis of the report
- II ☐ Priority
- III ☐ Non-establishment of report with regard to novelty, inventive step or industrial applicability
- IV ☐ Lack of unity of invention
- V ☒ Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
- VI ☐ Certain documents cited
- VII ☐ Certain defects in the international application
- VIII ☐ Certain observations on the international application

Date of submission of the demand 08 DECEMBER 1999	Date of completion of this report 27 JULY 2000
Name and mailing address of the IPEA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Joyce Bridgers ARUN CHAKRABARTI Telephone No. (703) 308-0196

Form PCT/IPEA/409 (cover sheet) (January 1994)★

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/US99/13295

I. Basis of the report

1. This report has been drawn on the basis of *(Substitute sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to the report since they do not contain amendments)*:

[X] the international application as originally filed.

[X] the description, pages 1-55 , as originally filed.
pages NONE , filed with the demand.
pages NONE , filed with the letter of ________.
pages ________ , filed with the letter of ________.

[X] the claims, Nos. 56-58 , as originally filed.
Nos. NONE , as amended under Article 19.
Nos. NONE , filed with the demand.
Nos. NONE , filed with the letter of ________.
Nos. ________ , filed with the letter of ________.

[X] the drawings, sheets/~~fig~~ 1-33 , as originally filed.
sheets/~~fig~~ NONE , filed with the demand.
sheets/~~fig~~ NONE , filed with the letter of ________.
sheets/~~fig~~ ________ , filed with the letter of ________.

2. The amendments have resulted in the cancellation of:

[X] the description, pages NONE .
[X] the claims, Nos. NONE .
[X] the drawings, sheets/~~fig~~ NONE .

3. [ ] This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the ~~Supplemental Box~~ Additional observations below (Rule 70.2(c)).

4. Additional observations, if necessary:

NONE

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/US99/13295

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. STATEMENT

Novelty (N)	Claims	3, 7-17	YES
	Claims	1,2, 4-6	NO
Inventive Step (IS)	Claims	7-17	YES
	Claims	1-6	NO
Industrial Applicability (IA)	Claims	1-17	YES
	Claims	None	NO

2. CITATIONS AND EXPLANATIONS

Claims 1, 2 and 4-6 lack novelty under PCT Article 33(2) as being anticipated by Mitra et al. (U.S. Patent 5,563,328) (08 October, 1996).

Mitra et al teaches a method for the cloning of intact, diversity-selected genes from within cassettes (Column 8, lines 35-40), the method comprising the steps of:

(a) identifying repeat DNA sequences which flank gene cassettes (Column 10, lines 49-50, Example 1, column 16, lines 44-46);

(b) hybridizing oligonucleotides to the repeat sequences which flank the gene cassettes and amplifying the sequences to provide DNA fragments which contain genes from within the cassettes (Column 10, lines 49-50, column 15, lines 38-39);

(c) ligating the DNA fragments into a vector (Column 10, lines 49-64, column 15, lines 42-44); and

(d) transforming the vector into an appropriate strain (column 10, lines 64-67, column 15, lines 49-53).

Mitra et al teaches the method wherein the diversity-selected genes are selected from toxin genes (Column 9, lines 3-6).

Mitra et al teaches the method wherein the diversity-selected genes comprise methyltransferase genes (Example 1, column 16, lines 34-63).

Mitra et al teaches the method wherein the oligonucleotides contain recognition sites which permit directional cloning (Column 17, lines 15-26).

Mitra et al teaches the method wherein the DNA fragments are ligated into the vector in an orientation that enables expression (column 15, lines 54-61).

Claim 3 lack an inventive step under PCT Article 33(3) as being obvious over Mitra et al (U.S. Patent 5,563,328) in view of Xu (U.S. Patent 5,492,823) (February 20, 1996). Mitra et al teaches the method of claims 1,2 and 4-6. Mitra et al does not teach the method wherein the diversity-selected genes comprise restriction endonuclease genes.
(Continued on Supplemental Sheet.)

Supplemental Box
(To be used when the space in any of the preceding boxes is not sufficient)

Continuation of: Boxes I - VIII

Sheet 10

V. 2. REASONED STATEMENTS - CITATIONS AND EXPLANATIONS (Continued):

Xu teaches a method wherein the diversity-selected genes comprise restriction endonuclease genes (Abstract and Example 1). It would have been obvious to substitute and combine the restriction endonuclease gene cloning method of Xu into the method of Mitra et al, since Xu states, "This method has been successfully employed to clone a number of genes coding for endonuclease including restriction endonuclease genes (Abstract, lines 4-7)". An ordinary practitioner would have been motivated to combine the method of Xu into the method of Mitra et al in order to achieve the express advantage of a method as noted by Xu which can be successfully employed to clone a number of genes coding for endonuclease including restriction endonuclease genes.

Claims 7-17 meet the criteria set out in PCT Article 33(2)-(3), because the prior art does not teach or fairly suggest the oligonucleotides having SEQ ID NOs: 5-91. Moreover, the prior art does not teach or fairly suggest screening the predetermined DNA sequence for TAACWA as well as CGTTRR.

Claims 1-17 meet the criteria set out in PCT Article 33(4) for industrial applicability.

------------------ NEW CITATIONS ------------------
NONE

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04

A1

(11) International Publication Number: WO 99/64632

(43) International Publication Date: 16 December 1999 (16.12.99)

(21) International Application Number: PCT/US99/13295

(22) International Filing Date: 11 June 1999 (11.06.99)

(30) Priority Data:

60/089,086	12 June 1998 (12.06.98)	US
60/089,101	12 June 1998 (12.06.98)	US

(81) Designated States: JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(71) Applicant (*for all designated States except US*): NEW ENGLAND BIOLABS, INC. [US/US]; 32 Tozer Road, Beverly, MA 01915 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): RALEIGH, Elisabeth, A. [US/US]; 32 Barton Street, Somerville, MA 02144 (US). VAISVILA, Romualdas [LT/LT]; 507 Sandy Bay Terrace, Rockport, MA 01966 (US). MORGAN, Richard, D. [US/US]; 31 Donovan's Way, Middleton, MA 01949 (US).

(74) Agent: WILLIAMS, Gregory, D.; New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915 (US).

(54) Title: RESTRICTION ENZYME GENE DISCOVERY METHOD

(57) Abstract

The invention is directed to direct cloning of intact genes, with a high probability that the orientation of expression is known in advance, and with a low probability of being associated with extraneous possibly toxic genes. The invention is particularly directed to obtaining genes encoded in DNA cassettes comprised of repeat sequences flanking variable open reading frames. The invention encompasses obtaining such cassette-encoded genes using oligonucleotides hybridizing to the repeated elements, cloning them and expressing them. Expression may employ tightly regulated vectors and useful strains disclosed. Methods for identifying restriction endonuclease and DNA methyltransferase genes in the absence of prior information about the sequences or biochemical specificities of these are also disclosed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

RESTRICTION ENZYME GENE DISCOVERY METHOD

RELATED APPLICATIONS

This Application is a PCT Application of U.S. Provisional Application Serial No. 60/089,101 filed 12 June 1998 and U.S. Provisional Application Serial No. 60/089,086 filed 12 June 1998, the disclosures of which are hereby incorporated by reference herein.

FIELD OF THE INVENTION

The invention is generally directed to the field of gene discovery, cloning and expression. A particular aspect of the invention is that it enables direct cloning of intact genes, with a high probability that the orientation of expression is known in advance, and with a low probability of being associated with extraneous possibly toxic genes

The invention is limited to genes of a particular kind, since some genes are more likely to be susceptible to cloning and discovery by this method than other genes. Accordingly, the invention is more specifically directed to cloning of genes found within arrays of gene cassettes separated by conserved repeated sequences. Based on present understanding, such arrays are found in prokaryotic organisms and contain genes that have functions that are selectively advantageous to a high level under certain circumstances but are not required under other conditions. Accordingly, some kinds of genes will not be found within these arrays, while other kinds of genes should be enriched in such arrays. Among the genes to be found in such cassette arrays are many genes of commercial interest. The kinds of genes of interest that may be expected in such arrays include:

Restriction enzymes, which are useful for a variety of procedures in molecular biology and which enable construction of may useful vectors.

Adhesins, which may allow a cell to attach to a particular surface. Enabling specific attachment to a particular surface rather than others has many uses in providing coatings and targeting molecules or organisms to locations of interest. Such adhesins may also mediate pathogenic processes when expressed by pathogenic organisms, and availability of an adhesin may enable competitive exclusion of such pathogenic organisms.

Small-molecule modifying enzymes, which may convert a toxic or other material abundant in a particular environment to another less toxic to humans or animals, or into a form more useful.

Specific toxin molecules that interact with a host organism, which may be useful for synthesis of inhibitors or antagonists of the toxin or for vaccine purposes.

Different examples of related cassette-encoded gene products will have common general properties (adhesins stick to things) but highly variable specificities (there are many different kinds of specific surfaces to stick to, from rocks to intestinal mucosa to urinary epithelium). Genes of this kind will be referred to below as "diversity-selected genes". The list of gene types above is not exhaustive.

BACKGROUND OF THE INVENTION

Hypervariable gene regions in prokaryotic organisms

Hypervariable regions, which show a high level of sequence divergence between closely related strains of the same species, are found at various positions in prokaryotic chromosomes. In some cases, genes present in one strain are absent entirely from a close relative. Examples of this phenomenon include so-called "pathogenicity islands", chromosomal elements that carry genes required for pathogenesis (McDaniel, et al., *Proc. Natl Acad. Sci. USA* 92(5):1664-1668 (1995)). Restriction enzyme genes are sometimes found in regions that are hypervariable in this way (Daniel, et al., *J. Bacteriol.* 170:1775-1782 (1988); Raleigh, *Mol. Microbiol.* 6:1079-1086 (1992); Barcus, et al., *Genetics*, 140:1187-

1197 (1995)). The mechanism of assembly and variation of these regions may depend on novel genetic mechanisms.

Integrons and superintegrons as hypervariable gene regions: mobile gene cassettes

Integrons (Hall and Collis, *Mol. Microbiol.*, 15(4):593-600 (1995)) are arrays of promoterless gene cassettes, separated by related DNA elements ("59 bp elements") that are sites of action for site-specific integrases related to the lambda integrase (Fig. 1). Each integron has at the 5' end a gene for the relevant integrase. Within the integrase gene is a promoter oriented toward the cassettes, upon which expression of all cassette-borne genes is dependent. Cassettes can be found as extrachromosomal nonreplicating circles, and these can be inserted into the array by the integrase. Characterized integrons are plasmid-borne, and the cassettes specify resistance to drugs or other toxic products (such as mercury). Ordinary integrons are small: up to 8 cassettes have been identified in one ordinary integron, and most have between one and three. It is thought that all the genes are expressed from the single promoter found within the sequence of the flanking integrase (Levesque, et al., *Gene* 142(1):49-54 (1994); Recchia and Hall, *Mol. Microbiol.*, 15(1):179-187 (1995)) (Fig. 1); in any event, promoter-like sequences are usually not identified within the gene cassettes. The plasmid location and the multiple-drug resistant character of integrons probably reflect the historical origins of the studies involved: they were found as a result of studies on horizontal transmission of drug resistance in bacteria isolated from clinical settings, where such behavior is selectively advantageous.

A superintegron (Mazel, et al., *Science*, 280(5363):605-608 (1998)) was recently described as a chromosomal array of a large number of gene cassettes mobilizable by a site-specific integrase obtained from an integron. This large array, found in *Vibrio cholerae*, may contain up to a hundred cassettes and may account for as much as 10% of the chromosomE (Barker, et al., *J. Bacteriol*, 176(17) 5450-5458 (1994)). The Manning laboratory identified this array in the course of studying a pathogenesis-related hemagglutinin (Franzon, et al., *Infect. Immun.*, 61(7):3032-3037 (1993)). Open reading frames within this array are separated by repeated sequences called VCR (for Vibrio cholerae repeats). These repeats are similar to but not the same as the "59 bp elements" of drug-resistance

integrons (Mazel, et al., *supra* (1998)). Manning's laboratory claims to have identified an integrase associated with *Vibrio cholerae* (Clark, et al., *Mol. Microbiol.*, 26(5):1137-1138 (1997)), and the Davies laboratory has published a description of such a gene from *Vibrio cholerae* (Mazel, et al., *supra* (1998)).

This superintegron is distinguished from the ordinary integrons in four respects: size, placement of promoters, replicon location, and the nature of the genes found within cassettes. In contrast to the best-studied integron examples, there appear to be 60 to 100 cassettes within the *V. cholerae* array; and since they are not all oriented in the same direction (Fig. 2), they cannot be expressed from a common promoter. Moreover, the functions encoded by the superintegron are apparently diverse, and some are possibly related to pathogenesis (Mazel, et al., *supra* (1998)). Some of the cassette-borne genes were related to some plasmid-encoded proteins (from database-matching of ORFs 3.1 and 3.2 of the sequence reported in (Barker, et al., *supra* (1994)), one was a heat-stable toxin (Ogawa and Takeda, *Microbiol. Immunol.* 37(8):607-616 (1993)), and one was similar to a lipoprotein gene (*vlpA*; from database matching of ORF2). Accordingly, we surmise (following Mazel et al) that this array may function to cluster genes related to pathogenicity and to the entrap genes specifying other biochemical functions.

Repeated sequences between gene cassettes in integrons and superintegrons

The sequences interspersed between gene cassettes are thought to be responsible for acquisition and exchange of gene cassettes among the various replicons on which they are located. These sequences, designated "59 bp elements" or "VCR elements" are diverse in sequence but display some common features. A consensus sequence was initially deduced for conventional "59 bp elements" (Hall, et al., *Mol. Microbiol.*, 5(8):1941-1959 (1991)), consisting of:

5' GYCTAACAA-TTCGTTCAAGCCGACGCCGC-T...
ICS

...-TC-GCGGC-GCGGCTTAACTC-ARGCGTTAGRY 3' (SEQ ID NO:92)
CS

It was later found that the relevant sequences varied in length and sequence within the segments (Hall and Collis, *supra* (1995)). Two most conserved segments could always be identified: 5' to a gene cassette (and at the 3' end of the sequence above; underlined) is found the "Core Sequence" (CS), GTTRRRY (SEQ ID NO:93); and 3' to a cassette (and at the 5' end of the sequence above; underlined) is found the "Inverse Core Sequence" (ICS), RYYYAAC (SEQ ID NO:94). These two elements are related as inverted repeats. Upon excision, the part of the sequence included in the extrachromosomal circle includes the sequence 3' to the gene as far at the G in the Core Sequence; the circle is completed with the remainder of the CS from the 5' end of the gene (TTAGRY (SEQ ID NO:95)).

The VCR elements were originally said to be unrelated to any other sequence (Barker, et al., *supra* (1994)) but were subsequently shown to conform with the specifications of the "59 bp elements" except for greater length (Mazel, et al., *supra* (1998); Clark, et al., *supra* (1997)): they consist of 124-bp direct repeats of imperfect dyad symmetry, and carry ICS and CS motifs at the ends. VCR elements were found nine times in the original sequence surrounding the putative hemagglutinin gene (Barker, et al., *supra* (1994)).

PCR has been used for characterization of integrons. Some studies employed primers annealing to the conserved integrase genes, or to *sulI,* a conserved gene found at the 3' end of many integrons (e.g.(Levesque, et al., *Antimicrob. Agents Chemother.*, 39(1):185-191 (1995); Sallen, et al., *Microb. Drug Resist.*, 1(3):195-202 (1995); Sandvang, et al., *FEMS Microbiol. Lett.*, 160(1):37-41 (1998)). Other studies have employed primers annealing to particular cassette-encoded genes (e.g. (Senda, et al., *J. Clin. Microbiol.* 35(12):2909-2913 (1996); Tosini, et al., *Antimicrob. Agents Chemother.*, 42(12):3053-3058 (1998)). However, it has been considered unlikely that these repeat sequences would enable acquisition of cassette-encoded genes by PCR, because of the degeneracy of the sequences and the secondary structure encoded by them (Hall and Stokes, *Genetica*, 90(2-3):115-132 (1993)). Mazel et al (supra, (1998)) were able to obtain cassettes by PCR using primers annealing to the VCR elements, however.

Background of restriction enzyme gene discovery

Restriction enzyme properties.

Restriction enzymes are the workhorses of molecular biology research. They specifically recognize sites in DNA of 4 to 8 basepairs in length, with extremely high selectivity--that is, a site with one mismatch is typically recognized with an affinity one-thousandfold less than the affinity shown for the correct site. This high degree of selectivity is essential for use in practical applications.

Known restriction enzymes recognize over 200 different specific DNA sequences (Roberts and Macelis, *Nucleic Acids Res.*, 26(1):338-350 1998)) and many of these are commercially available. However, the potential number of different sites is much larger: 32,512 distinct 8-base sites might be recognized [((4^8/2)-256): a site 8 bases in length with four possible bases at each position; which can be recognized in either of two complementary strands; minus 256, since 8-base palindromes each read the same in the two strands].

Enzymes with 8 bp recognition sites (8-cutters, such as NotI, SfiI, SwaI, PacI and PmeI) are of particular utility. These enzymes are used for constructing maps of and manipulating DNA from high-complexity sources, such as the genomes of humans and other higher eukaryotes. This utility arises from the rarity of the sites (once per 65,000 bp for palindromic sites), enabling for example the isolation of a whole gene with large introns on a single DNA fragment.

Of the twelve known specificities with 8 bp recognition sites, two were found in *Pseudomonas spp.* nine in *Streptomyces* or other high G+C gram positive bacteria, and one in *Staphylococcus*. Sequence information is available for six of these, the two *Pseudomonas* isolates and four from high G+C organisms.

Competing approaches to restriction enzyme discovery.

In the past, two broad approaches have been taken to the problem of finding new restriction enzymes: screening for new enzymatic activities, and changing existing enzymes to recognize new sites.

1) Screening of crude extracts of individual prokaryotic strains (obtained from strain collections or natural environments). A test substrate (e.g. phage lambda DNA) is incubated with such an extract, and the digest visualized by agarose gel electrophoresis. This standard approach identifies at least one site-specific nuclease in about 25% of crude extracts screened, with the routine use of targets of combined complexity of about 200 kb.

This approach has two critical defects. First, the fraction of such enzymes recognizing new sites is now very low. In part this may be due to its bias toward identifying enzymes with recognition sites between four and six bp in length and inefficiency in detecting enzymes with larger targets, which are frequently not present in the target substrates.

The second defect is that is extremely labor-intensive. Each strain must be examined individually, and several of the steps involved are projects in themselves: culture growth, cell lysis, and extract clarification each can be a custom procedure. The quality of crude extract preparations varies greatly among isolates, in the extent of contamination with extraneous nucleases, DNA binding proteins and proteases.

In the specific case of *Pseudomonas* and its relatives, extracts are frequently difficult to handle due to extensive nuclease contamination. *Xanthomonas* strains (which are relatives of *Pseudomonas*) frequently give cultures that are hard to collect by centrifugation due to copious extracellular polysaccharide production, and extracts are difficult to clarify for the same reason.

2) Mutational alteration of existing enzymes so that they recognize new sequences. Starting with enzymes recognizing 6 base pairs for which structural information is available, attempts have been made to alter specificity by site-directed, random or random cassette mutagenesis (e.g. (Dorner and Schildkraut, *Nucleic Acids Res.* 22(6):1068-1074 (1994); Heitman and Model, *EMBO J.* 9(10):3369-3378 (1990); Ivanenko, et al., *Biol. Chem.* 379(4-5):459-465 (1998); Hager, et al., *J. Biol. Chem.* 265(35):21520-21526 (1990) and I. Schildkraut, personal communication). Although this work may eventually yield useful products, it has not yet produced an increased specificity (recognizing more bases) or altered specificity (recognizing a different sequence of the same length).

Background of restriction enzyme gene clone identification and cloning

Restriction enzymes are found in a wide variety of prokaryotic organisms, many of them with fastidious growth requirements and frequently in low amounts. For purposes of commercial production, it is most useful to be able to produce a restriction enzyme in a well-understood and genetically tractable bacterial host such as *Escherichia coli*. The many tools for gene expression and regulation, as well as for genetic manipulation of the host cell, enable preparations to be made with higher purity and lower cost. Accordingly it is very useful to obtain the genes for endonucleases as molecular clones.

Methyltransferase selection method

One method for identifying the presence of a restriction enzyme gene in a clone library is to rely on the presence and expression of a closely-linked gene for a cognate DNA methyltransferase (Wilson, U.S. Patent No. 5,200,333 (1993)). Such methyltransferase enzymes recognize specific DNA sequences and add a methyl group to an A or C residue within the sequence. This modification prevents cleavage by the endonuclease, thereby protecting the host genome from lethal damage. If such a methyltransferase gene is present in a clone library and effectively expressed, the DNA of that clone will be protected from digestion. This enables selection for the clone in vitro: plasmid clone DNA is purified from a pool of clones and digested with the desired endonuclease enzyme. The methyltransferase clone will not be digested, while other clones in the library, (which are found in different cells) will be destroyed. Retransformation following such a procedure allows establishment of a selected pool, in which representation of the methyltransferase gene is greatly enriched. If the endonuclease gene is adjacent to the methyltransferase gene, as is often the case, then that gene (or a portion of it) will also be recovered frequently. This method is called the "methyltranferase selection" method. It is quite useful when three conditions obtain: a cognate methyltransferase exists; the genes for the two functions are tightly linked in the DNA; and the methyltransferase is expressed in *E. coli*.

Several modifications have been added to this basic method, enabling isolation of the endonuclease gene when the first clone does not contain the complete endonuclease gene or when the methyltranferase must be expressed in the cell first, before the endonuclease can be introduced (the "two-step" method) (Brooks and Howard, U.S. Patent No.5,320,957 (1994)).

Degenerate methyltransferase-motif PCR method

A second method for identifying the presence of a restriction system gene pair in a clone library is to rely on the presence of conserved polypeptide motif elements found in the DNA methyltransferase proteins (Klimasauskas, et al., *Nucleic Acids Res.* 17:9823-9832 (1989); Lauster, et al., *J. Mol. Biol.*, 206:305-312 (1989); Posfai, et al., *Gene* 74(1):261-265 (1988)). This method is most useful when three conditions obtain: a cognate methyltransferase exists, the genes for the two functions are tightly linked in the DNA, and the methyltransferase is not effectively expressed in *E. coli*. Because the methyltransferase is not effectively expressed, the methyltransferase selection method cannot be used. Briefly, this alternative method is as follows: the polypeptide sequence of the conserved polypeptide motif elements is reverse-translated into a pool of DNA sequences each capable of specifying the polypeptide sequence in question. This pool is called a degenerate pool, because the genetic code is degenerate--several different DNA triplets can specify the same amino acid in many cases. This degenerate pool of oligonucleotides is then used to amplify fragments of DNA from genomic DNA or from a clone library. The sequence of the PCR fragments is then determined, enabling design of further non-degenerate (unique) primers that detect the presence of the proper sequence in the genomic DNA or the clone library by hybridization or PCR. Adjacent DNA sequence can then be obtained by the inverse-PCR method or by Southern blot screening procedures; further sequence can be determined; and finally the complete restriction system can be assembled. This method can be used either alone or in combination with other procedures (below) to isolate the methyltransferase gene and the adjacent endonuclease gene.

"Methylase indicator" DNA damage method.

Another method for identifying clones containing methyltransferase genes (Piekarowicz, et al., *J. Bacteriol.* 173:150-155 (1991); Piekarowicz, et al., *Nucleic Acids Res.*, 19:1831-1835 (1991); Piekarowicz and Weglenska *Acta Microbiol. Po..*, 43(2):229-231 (1994)) relies on methylation-dependent restriction systems McrA, McrBC and Mrr (Heitman and Model, *J. Bacteriol.* 169&7):3243-3250 (1987); Heitman and Model, *Gene* 103:1-9 (1991); Waite-Rees, et al., *J. Bacteriol.*, 173(16):52-7-5219 (1991); Raleigh and Wilson, *Proc. Natl. Acad. Sci. USA* 83:9070-9074 (1986); Kelleher and Raleigh, *J. Bacteriol.*, 173(16):5220-5223 (1991)) and on the *dinD1::lacZ* operon fusion, to enable a method to screen for clones that contain methyltransferase genes. Strains with temperature sensitive mutations in *mcrA, mcrBC,* and *mrr* are permissive at high temperature for expression of methyltransferase activity by cloned foreign genes. When these restriction functions are active however (at low temperature), they will cleave DNA methylated by foreign methyltransferase enzymes. This cleavage leads to generation of a signal that induces expression of the endogenous DNA damage inducible (SOS) regulon. The *dinD1::lacZ* transcriptional fusion between one of the genes in this regulon (*dinD*) and the *lacZ* gene is then induced, and β-galactosidase is expressed. Action of the β-galactosidase allows the colonies turn blue on plates containing Xgal. Thus, colonies from a clone library that are white (or light blue) at high temperature but dark blue at low temperature are methyltransferase clone candidates.

N-terminal sequence/degenerate PCR method

It may occur that a methyltransferase gene cannot be identified, or that a methyltransferase gene can be identified but the open reading frame specifying the endonuclease is uncertain. In these cases, an additional useful procedure for identifying the gene for the endonuclease specifically can be applied when the endonuclease can be purified in sufficient quantity and purity from the original organism. In this method, the endonuclease polypeptide is purified to homogeneity and subjected to N-terminal polypeptide sequencing. The polypeptide sequence is reverse-translated into a pool of DNA primers capable of specifying the appropriate sequence, and these primers are used to amplify a

portion of the endonuclease gene from genomic DNA of the original organism or from a clone library.

This procedure can be used alone to obtain a portion of an endonuclease gene, or in combination with other methods, such as the degenerate methyltransferase-motif PCR method (Morgan, U.S. Patent No. 5,543,308 (1996)) to obtain portions of genes for both components of the restriction system. The complete genes can be assembled with the assistance of Southern blot or by further direct or inverse PCR methods. If the cognate methyltransferase gene cannot be obtained or cannot be expressed, the stability and utility of solo endonuclease clones will be severely compromised. Such clones can be stabilized with the use of heterospecific methyltransferase genes, which were not associated with the endonuclease in the original host, if they recognize the same or a related sequence and prevent the endonuclease from cleaving its recognition sequence (Wilson and Meda, U.S. Patent No. 5,246,845 (1993)).

Endo-blue method

Another method for identifying the presence of an endonuclease gene in a clone library, independently of the presence of the cognate methyltransferase gene, is to introduce the library into a restrictionless host *E. coli* strain containing a reporter of DNA damage. This method is related to "methylase indicator method" above, but the strain used contains no restriction activity specific for methylated DNA. In this case, cleavage occurs due to expression of the restriction enzyme, thereby inducing the SOS regulon (and the *dinD::lacZ* indicator) directly rather than through the action of the methyltransferase and endogenous restriction activities. Action of the β-galactosidase then allows the colonies to turn blue on plates containing Xgal.

This indicator can be used to identify restriction endonuclease clones when a modification methyltransferase gene is poorly expressed, so that some DNA damage occurs despite its presence, or without the methyltransferase when conditional activity of the endonuclease can be obtained. For example, the endonuclease in question may be inactive at low growth temperatures but somewhat active at higher growth temperatures. The latter situation obtains, for example, with some restriction endonucleases originally expressed in

hyperthermophilic organisms, which normally grow at very high temperatures (Fomenkov, et al., U.S. Patent No. 5,498,535 (1996); Fomenkov, et al., *Nucleic Acids Res.* 22(12:2399-2403 (1994)).

Background of regulation of gene expression in cloned genes.

Regulation of expression from vector promoters

In very many instances the problem for the experimenter is to obtain sufficient expression from cloned DNA to enable useful amounts of a gene product to be made in the new cellular environment. Accordingly, there are many expression vectors available that provide one or more promoters enabling high-level transcription activity proceeding through the location at which foreign DNA is to be introduced. Frequently these vectors are provided with a gene for a regulatory molecule such as a repressor of transcription able to regulate expression from the promoter provided, or are used in host organisms that themselves provide such a regulator. In this way, the expression desired can be provided on demand, ie. during induction of specific expression. Many such vectors are described in the art (Sambrook, et al., Molecular Cloning: A Laboratory Manual (1989)).

In some instances, the reverse problem occurs: the product expressed from the cloned DNA is toxic to the cell expressing it for some reason, and ordinary vectors designed for expression at high levels express too much of the toxic product, even in the absence of specific induction. Accordingly, vectors have been described that are designed to express cloned genes at extremely low levels in the absence of induction. The best known of these is the T7 RNA polymerase-dependent expression system designed for use in *E. coli* (Studier, et al., *Meth. Enzymol.,* 185:60-89 1990)). In this system, cloned genes are expressed from a promoter of transcription that is not recognized at all by any endogenous *E. coli* RNA polymerase holoenzyme. Instead, the promoter employed is recognized by the RNA polymerase of bacteriophage T7. This polymerase is not encoded in the *E. coli* genome. This system enables the construction of a clone with toxic properties in the absence of the required RNA polymerase. The clone can then be introduced into a suitable strain into which the T7 RNA polymerase gene has been introduced previously, or the polymerase gene can be introduced by infection with a phage-borne clone.

Inhibition of expression from indigenous promoter-like sequences

An additional problem with toxic proteins can be encountered when the foreign DNA, introduced into the expression vector, itself contains sequences recognized by the *E. coli* expression apparatus. The specific regulators provided by the vector/host combination will not regulate promoter activity originating within the cloned sequence. In some cases this expression may be the result of specific promoter recognition, but it may also arise simply from adventitious promoter-like activity in DNA, particularly in DNA rich in A+T (Miller and Simons, *Mol. Microbiol.*, 4(6):881-893 (1990)). In such instances a useful method of control is to provide, in the vector, a regulatable promoter opposing the direction of translation of the cloned DNA (Cole and Honore, *Mol. Microbiol.* 3(6):715-722 (1989); Adhya and Gottesman, *Cell* 29(3):939-944 (1982); Elledge, et al., *Proc. Natl. Acad. Sci. USA*, 86(10):3689-3693 (1989); Simons, and Kleckner, *Annu. Rev. Genet.*, 22:567-600 (1988); Roberts, et al., International Publication No. WO 99/11821 (1999)). A high level of transcription in the direction opposite that needed for polypeptide expression can interfere with expression in at least two ways. First, it can occlude transcription in the direction needed for expression; and second, it can prevent translation by allowing formation of RNA-RNA hybrids between the RNA used for expression of the toxic protein and the RNA directed in the opposite sense (antisense RNA).

Cloning into an expression vector for tight regulation

Restriction endonucleases, which cleave DNA at particular sequences, are normally associated with protective modification methyltransferases. In the present method it is quite likely that the gene for such an endonuclease will be isolated without its partner methyltransferase gene. Very tight regulation of the cassettes thus cloned is therefore critical.

A convenient tightly regulated expression plasmid, pLT7K, is available into which pooled PCR fragments can be cloned (Roberts, *supra* (1999)). In this vector, two levels of control are available: expression is inducible and inhibition is repressible. A T7 gene 10 promoter reads into one side of the cloning site; LacI provided by the vector represses expression from this promoter, as is expression

of the T7 RNA polymerase provided by the host cells used for expression. Further control can be obtained by the use of pLysP, which expresses an inhibitor of T7 RNA polymerase.

To further reduce expression directed by the cloned fragment, and residual leaky expression from the T7 promoter, the λ pL promoter reads into the other side of the cloning site, antagonizing expression from pT7. This antagonistic transcription is regulated by λ cI^{857}, a thermosensitive repressor. At 40°C and in the absence of IPTG therefore, essentially no expression was observed; at 30°C, some leaky expression is seen; at 30°C in the presence of IPTG, moderate levels of expression can be achieved. This vector has successfully been used to establish the *pacIR* and *nlaIIIR* genes (encoding the restriction enzymes PacI and NlaIII) in the absence of methyltransferase protection, and to express the genes.

SUMMARY OF THE INVENTION

A general object of the invention is to provide a procedure for obtaining clones of diversity-selected genes. A specific object of the invention is to provide a method for identifying a repeat sequence suitable for identification and cloning of gene cassettes found in arrays and separated by repeat sequences. A specific example of such a repeat sequence family with 74 members is provided together with the sequences of four contiguous DNA stretches comprising one or more cassette arrays. A further specific object of this invention is to provide a procedure for cloning cassettes from such arrays, by PCR directed by oligonucleotides hybridizing with the repeated sequences flanking the cassettes. A specific example of such a PCR procedure is provided. A further specific object of this invention is to provide a procedure for cloning such PCR fragments into an expression vector able to stabilize toxic genes such as restriction enzymes. A specific example of such a gene clonable by this procedure is provided. A further specific object of the invention is to provide a means of identifying particular cloned genes of interest. Accordingly, three methods of identification are provided: one method relies on identification by means of protein sequence similarity; a second method relies on an indirect report of gene activity; a third method relies on direct test of biochemical properties. In accordance with this method, two novel strains that enable provision of indirect report of expressible cloned nuclease genes in the context of the vector

pLT7K are provided, together with a method of use. A further specific object of the invention is to provide a method for obtaining expression clones of active restriction enzyme genes without prior knowledge of their biochemical activity or DNA sequence. A specific example of a procedure for obtaining such a clone is provided.

Since the invention relates to genes found in a particular sort of hypervariable locus, a description of what sorts of genes these will be is provided.

Features of gene cassettes useful for cloning methods.

In the particular case of hypervariable loci that are integrons or superintegrons, these regions provide a mechanism for discovery of diversity-selected genes. The features of these systems enable isolation of DNA enriched for certain kinds of genes including restriction enzyme genes, and also enable the cloning, sequencing and expression of products encoded in this DNA.

Three features of cassette arrays are particularly useful for cloning purposes:

- Each gene (rarely, a pair of genes) is embedded in a predictable sequence context--a particular kind of repeated DNA sequence is found on each side.

- Most genes found such arrays are in the same orientation relative to the flanking sequences.

- Expression of cassette-encoded genes is frequently directed from outside the cassette.

These properties make it likely that genes cloned by PCR from the flanking repeat elements will be intact, will be in an orientation specified in advance relative to the cloning vehicle, and can be regulated by expression signals in the cloning vehicle. This yields a set of DNA fragments in which each gene (rarely, a pair of genes) is embedded in a manipulable sequence context--suitable sites for cloning can be included at the 5' ends of the PCR primers.

A difficulty with these repeat sequences is that the members of the repeated array are degenerate, so that PCR primers hybridizing to most or all of the members of the array are difficult to design. Accordingly it is important to have available a large number of such sequences, enabling design of multiple family-specific primers. Such a collection of repeat sequences is identified and characterized in accordance with this invention.

A second difficulty with these repeat sequences is that individual members of the repeated array display imperfect dyad symmetry elements, making it likely that PCR primers designed will form hairpins or primer dimers and so fail to prime DNA amplification. Accordingly, it is important to design primer that anneal to portions of the repeats that do not display these features. Primers that are able to hybridize with or that enable amplification from many cassettes are provided in accordance with this invention.

Expression cloning of cassette-encoded genes.

A very large number of uncharacterized cassettes may potentially be obtained by this method, so that the experimenter will require some procedure for sorting through these for functions of interest. Accordingly, the present invention provides a method for obtaining expression of cassette-encoded functions even when toxic, by cloning these into an appropriate vector, such as the pLT7K vector described in International Publication No. WO 99/11821 (Roberts, et al., (1999)).

This vector has the advantage (in addition to those provided in the original patent) that it can be used in two configurations in this application. Depending on the orientation of cloning sites on the PCR primers, the expression condition can be either 30 C + IPTG or 40 C - IPTG; and the repressed condition suitably the reverse. This enables flexibility in screening or selecting for molecules that display activity sensitive to temperature, and in selecting storage conditions for the clone library obtained.

Strain enabling indirect report of nuclease activity.

A test of function is provided that enables detection of a minority of expression clones of interest in the context of the T7-RNAP dependent regulation required by the vector pLT7K. This test detects nuclease or other DNA damaging activity by SOS induction of *dinD::lacZ* alleles. Two strains are provided:

ER2745: (F⁻ λ⁻ *fhuA2 [lon] [dcm] ompT lacZ::T7 gene1 gal sulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10*--TetS)*2 R(zgb-210::Tn10* --TetS) *endA1) dinD2::MudI1734 (KanR, lacZ*$^{+}$)

ER2746: (F⁻ λ⁻ *fhuA2 glnV44 e14- rfbD1? relA1? endA1 spoT1? thi-1 Δ(mcrC-mrr)114::IS10 lacZ::T7 gene1 dinD2::MudI1734 (KanR, lacZ(ts))*

The former can be used at either 30°C or 42°C to indicate DNA damage with a dark blue color against a background of lighter blue colonies. The latter can be used at 30°C up to and including 37°C to indicate DNA damage with blue color of any shade against a background of white colonies. Accordingly, libraries of cassettes cloned into pLT7K (or a derivative) in an orientation such that expression is driven by pT7 in the presence of T7 RNAP and inhibited by expression from λ pL can be screened for activity at 30°C or 37°C (with or without the presence of IPTG) in either strain. Libraries of cassettes cloned into pLT7K in an orientation such that cassette expression is driven by λ pL and inhibited by pT7 can be screened at 37°C (with or without IPTG) in either strain or 40°C (with or without IPTG) in ER2745 but not ER2746. In each case the presence of activity is indicated when a colony turns bluer than the majority class, and when this property is stable upon reisolation as a single-colony derivative of the original transformant.

These strains may similarly be used to indicate DNA damage provoked by any agent, including enzymes that are not nucleases, by chemical agents, or by radiation. These strains are most distinctively useful when the damage produced results pursuant to a regulated change in the state of T7 RNA polymerase expression as provided within these strains.

Kinds of genes for which this method may be applied.

In accordance with this invention, a limitation is provided for the kinds of genes for which the invention is useful. Some kinds of genes are likely to be present in cassette arrays, while others are unlikely to be present in them. The original cassettes of known function all specified resistance to drugs or other antibacterials. There is no a priori reason to suppose that integrons cannot mediate the spread of functions other than drug resistances. Types of genes likely to be enriched in such arrays include functions useful individually or in pairs, and subject to highly variable selective value. Typically such genes will be subject to strong episodic selection, very important some of the time but not useful at all the rest of the time. In some cases they will be episodically essential--necessary for cell survival: drug resistance factors, restriction-modification systems. In other cases they may be episodically of very high selective value, but not necessary for survival as such. Examples would include specific adhesins that allow the cell to attach to a particular surface in a rich environment; specific enzymes that modify an abundant material in the cellular environment to convert it to a form usable as nutrition; or specific toxin molecules that interact with a host organism. Many individual members of a particular species will elaborate gene products that have common general properties (adhesins stick to things). An important feature of relevant gene products, however, is that among the population will be found examples with highly variable specificities (there are many different kinds of specific surfaces to stick to, from rocks to intestinal mucosa to urinary epithelium).

Cassette arrays therefore will be enriched for genes that are subject to selection for diversity as described above: that is, genes that are advantageous when rare but of no particular use when frequent in the population; and those episodically required.

Types of genes expected to be absent from such arrays include all of the basic components of the cellular maintenance machinery: DNA replicases, basic transcription factors such as vegetative RNA polymerase, the translational machinery, enzymes of small molecule metabolism central to cellular physiology such as those of the tricarboxylic acid cycle. They should be absent for two

reasons. First, no selective advantage is expected from maintaining variability as such in the pool of alleles available to a population of cells. Second, many such proteins must maintain (conserve) specific interactions among several different proteins (replicase/RNA polymerase/translation initiation factor interactions for example).

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a schematic of the structure of characterized integrons, arrays of gene cassettes (thin lines; fn1, fn2, fn3) separated by repeated sequences (filled boxes; 59 bp elements). These are assembled by the action of a site-specific integrase (large box; intI) by insertion into attI (arrows) of extrachromosomal circles (cassette). Cassettes are transcribed from a promoter within the integrase gene (arrow). Many integrons are associated with a conserved sulfonamide resistance gene (sulI) that is not part of the integron itself.

Figure 2 is a schematic diagram of a fragment of a superintegron identified in Vibrio cholerae. Open reading frames (1-9 and mrhA, mrhB) are separated by repeats (boxes) that are similar to 59 bp elements of integrons

Figure 3A-3E is an alignment of some of the PAR elements (SEQ ID NO: 96 through SEQ ID NO:116), those identified in superintegron contig 1 (SEQ ID NO:1) by the motif search procedure described in Example 1. Consensus lines show bases shared by all (top line), 90% (second line) or the majority (third line) of the elements in the alignment. Individual entries are the same as the majority consensus except for the bp shown.

Figure 4. is a dotplot display illustrating an alternative method for identifying repeated sequences.

Figure 5. illustrates the self-complementarity of an individual PAR element (SQUIGGLE display of the output of FOLD in the GCG program set).

Figure 6 illustrates alignments of subfamilies identifiable in the set of PAR elements herein (SEQ ID NO:5 through SEQ ID NO:78) shown in Table 1. Panels

A-D, families 1-4. Each family alignment includes PAR2 as an outgroup member, since PAR2 is the most distantly related of the elements identified. Families were identified as bushy groups in a phylogenetic tree generated from the CLUSTAL alignment of the 74 elements.

Figure 7 illustrates the location of oligonucleotides used for Southern blots (panel A) and PCR fingerprinting (Panel B) in relation to the majority consensus of all PAR elements and in relation to a typical cassette.

Figure 8 illustrates a Southern blot hybridization of a mixture of Oligonucleotides 2-5 (SEQ ID NO:79 through SEQ ID NO:83; Fig 7, see also Table 2) to *P. alcaligenes* DNA.

Figure 9 displays an agarose gel of PCR products generated from chromosomal DNA of isolates of six *Pseudomonas species* by the use of oligonucleotides 6 and 7 illustrated in Fig. 7.

Figure 10 illustrates the scheme for forming a clone library of cassette-encoded open reading frames and expression of their products from pLT7K.

DETAILED DESCRIPTION OF THE INVENTION

In accordance with one embodiment of the invention, there is provided a novel method for the direct cloning and expression of diversity-selected genes residing in cassette arrays. In general, the method comprises the following steps, although as the skilled artisan will appreciate, modifications to these steps may be made without adversely affecting the outcome:

1) The class of genes of interest is identified and the suitability of the class for the method is evaluated.

In one embodiment of the invention the desirable genes are those for restriction endonucleases and modification methyltransferases. Types of genes likely to be enriched in cassette arrays include functions useful to the organism individually or in pairs, and subject to highly variable selective value. A function may be identified as likely to be encoded by genes in such arrays when a survey of different isolates of a species determines that the presence of the function, or its

specificity, is variable within the collection of isolates. For example, a survey of isolates of Escherichia coli reveals that many isolates but not all isolates express type II restriction enzymes; and that of those that do, the specificity of the enzyme (the sequence recognized) is variable, with many different specificities determined within the species. Candidate functions that will be subject to such variation include, in addition to restriction enzymes, cell surface antigens such as polysaccharide antigens or polypeptide antigens or secreted molecules; adhesins of various sorts such as fimbrial proteins, pilus proteins or outer membrane proteins; transporters of small molecules, especially those with narrow specificity; exported functions such as toxins, hemolysins, hemagglutinins, kinases and signalling molecules; detoxifying enzymes such as drug resistance determinants; catabolic enzymes specific for compounds episodically available (excluding those required for central metabolic pathways such as the tricarboxylic acid cycle); enzymes for biosynthesis of rare sugars (excluding those required in all cells, such as ribose, deoxyribose, and sugars of the cell wall), especially of those sugars that form part of the pericellular envelope.

In one embodiment of the invention, the desirable genes are those for restriction endonucleases and modification methyltransferases. Typically such genes will be subject to strong episodic selection, very important some of the time but not useful at all the rest of the time. Restriction functions can provide a very powerful protection against the invasion of foreign DNA (as when a bacteriophage infects the cell). This protection will only be effective if the host from which the bacteriophage did not carry the same restriction functions--otherwise its DNA would already carry the protective modification pattern of the invaded cell. Populations should therefore carry a wide variety of specificities of restriction-modification systems, and should switch them rapidly on an evolutionary time-scale. In accordance with this expectation, many restriction systems are found on plasmids. Integron-like structures provide an easy way to acquire a restriction system from a foreign source such as a plasmid, which might not establish itself successfully. The existence of the repeat elements would also provide a mechanism for a high rate of loss (by unequal crossing-over or slipped-mispairing during replication), thereby conferring a high degree of fluidity upon the cell's complement of restriction-modification systems.

2) DNA preparation

Genomic DNA is prepared from a strain of interest or from a consortium of strains or from an environmental source by methods known in the art, or DNA of plasmid, cosmid, BAC or PAC clones of genomic DNA from such sources is prepared.

3) Suitability of the DNA preparation for use of the method.

This is evaluated by determining the presence of repeated sequence arrays. Preferred methods are Southern blot hybridization or PCR fingerprinting using hybridization probes or PCR primers listed in Example 1. Other suitable primer pairs may be designed based on sequences listed in Example 1, or on other particular repeat sequences identified by methods described in Example 1. A DNA preparation is suitable for use if a hybridization signal is obtained or PCR products are obtained or both. In a preferred embodiment, PCR conditions are optimized using a non-proofreading DNA polymerase, by varying primer-template ratio, annealing temperature, magnesium ion concentration and extension time.

4) Cassette isolation

The DNA preparation is subjected to PCR employing a pair of primers annealing to repeat sequences flanking the cassettes and containing at their 5' ends sites for endonucleases compatible with cloning into a plasmid vector. Preferred primer pairs include those listed in Example 2; other suitable primer pairs may be designed based on sequences listed in Example 1, or based on other particular repeat sequences identified in the literature or by methods described in Example 1. In a preferred method, PCR conditions are optimized using a proofreading DNA polymerase, by varying primer-template ratio, annealing temperature, magnesium ion concentration and extension time. PCR fragments are purified away from primers, for example by means of size fractionation using commercially available kits.

5) Cassette cloning

The PCR fragments are digested with the appropriate restriction endonucleases for cloning, in one preferred procedure with XhoI and XbaI. The digested fragments are ligated into a suitable vector. Preferred vectors for this purpose have two particular properties. First, they contain a cloning site disposed to allow directional cloning of fragments. Directional cloning methods include the process of digesting the vector with two different restriction enzymes such that the single-stranded extension at one end does not hybridize the single-stranded extension at the other end of the vector backbone containing the origin of replication; and then ligating, to that vector backbone, DNA fragments having an extension at one end that hybridizes with one single-stranded extension of the vector backbone, and having an extension at the other end that hybridizes with the other single-stranded extension of the vector backbone. Other directional cloning methods can be envisioned, including for example the use of site-specific recombination enzymes, or hybridization of extensions provided by methods other than restriction enzyme cleavage. Second, preferred vectors contain two independently regulatable expression signals, one on each side of the cloning site described above and directed toward expression of the sequence resident at the cloning site. One preferred vector is pLT7K (Roberts, et al., International Publication No. WO 99/11821 (1999)). Other vectors include pBR322, pUC19, pACYC184, pSC101, pBeloBAC11, or their derivatives.

6) Strain choice

The ligated products are transformed into a strain suitable for screening or selecting for cassettes encoding desirable functions. For this purpose the strain must be compatible with the expression regulation signals provided by the vector chosen and must enable the method to be used for identifying desired cassettes.

In the simplest case, sequencing large numbers of cloned cassettes and subsequently evaluating the sequence information will identify cassettes of interest by bioinformatic methods. Such methods include matching the cassette-encoded sequences against public or private databases by means of similarity-determining algorithms such as BLAST or FASTA, or by employing a motif or pattern-based

search of the cassette-encoded sequences employing databases such as the PROSITE profiles database or the BLOCKS and PRINTS databases (Patterson, M. and Handel, M. (1998) Trends Guide to Bioinformatics, Elsevier Science, Cambridge, UK). In this case there are few constraints on strain or vector choice.

In other cases, cassettes of interest will be identified by sequence-based methods such as PCR or hybridization with probes. In these cases there are also few constraints on strain or vector choice.

In a preferred embodiment, cassettes of interest will be identified by activity expressed in vivo. In this case the choice of strain and vector is constrained: vector and strain must be compatible, enabling suitable regulation of cassette expression; by the nature of the activity to be expressed will also constrain strain choice.

In one embodiment, the activities to be expressed are modification methyltransferase activity or restriction endonuclease activity, both of which are amenable to identification by indirect report of activity based on damage inflicted in intracellular DNA and induction of the DNA damage repair response. Two preferred strains ER2745 (F⁻ λ⁻ *fhuA2 [lon] [dcm] ompT lacZ::T7 gene1 gal sulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10*--TetS)2 *R(zgb-210::Tn10* --TetS*) endA1) dinD2::MudI1734 (KanR, lacZ*⁺). and ER2746: (F⁻ λ⁻ *fhuA2 glnV44 e14- rfbD1? relA1? endA1 spoT1? thi-1 Δ(mcrC-mrr)114::IS10 lacZ::T7 gene1 dinD2::MudI1734 (KanR, lacZ(ts))* are strains compatible with the vector pLT7K.

ER2745 is derived from the particular strain background normally used for T7 RNAP-directed expression, and is ultimately a derivative of *E. coli* B. The protein expression properties of this strain background are well understood. This strain is transformable with DNA, but the level of transformation obtained is less than with other strains. The amount of the indicator *lacZ* expressed in the absence of DNA damage is relatively high, leading to light-blue colonies on Xgal plates even when no damage has occurred.

ER2746 carries a thermosensitive *lacZ* moiety. This is useful because it lowers the light-blue background color observed on X-gal by the original *dinD* indicator allele. Discrimination between clones inducing some damage (which are

of interest) and those inducing no damage (which are not) is improved in this situation. However, this allele cannot be used to detect DNA damage at high temperature (>37°C), because the *lacZ* moiety of the indicator fusion is inactive, and will remain white even in the presence of extensive DNA damage. This was demonstrated by testing at various temperatures for induction of blue color by nalidixic acid, a well-characterized DNA damaging agent, on plates containing X-gal.

Further refinement of this system is possible; for example, transcriptional fusion of a drug-resistance gene to a damage-inducible promoter should allow selective isolation of clones of interest, rather than the more-laborious screening procedure. Use of a variety of drug concentrations would then allow isolation of clones with different levels of DNA-damaging activity. Introduction of a *recD* mutation would inactivate the major ATP-dependent double-strand exonuclease of the cell, while an *xth* mutation would inactivate ExoIII, the major ATP-independent double-strand exonuclease. A triply nuclease-deficient strain should be viable but may not stably maintain the plasmid (Niki, et al., *Mol. Gen. Genet.* 224(1):1-9 (1990)).

Other DNA damage-inducing promoters that can be used include those identified by (Lewis, et al., *J. Bacteriol.*, 174:3377-3385 (1992); Lewis, *J. Mol. Biol.*, 241:506-523 (1994)): these are promoters of *recA, lexA, uvrA, uvrB, dinG, polB, uvrD, ruvAB, umuDC, sulA, dinH, dinI, sosA, sosB, sosC, sosD.* Other SOS-inducible genes identified include *recN, dinB* and *dinF* (Walker, *Microbiological Review*, 48:60-93 (1984)). Some other indicator/reporter genes that can be used were reported in (Fomenkov, et al., *supra* (1995).

7) Cassette identification: endonuclease genes

Following transformation or electroporation of the cassettes ligated with the chosen vector into the chosen strain, transformants are plated onto suitable media. In the preferred procedure, the vector is pLT7K, the strain is ER2746, plates are Luria-Bertani plates with ampicillin, and incubation is at 40°C. Colonies are replica plated onto plates containing Xgal with or without IPTG (at concentrations varying from 0.1 mM to 1 mM) and one set of replicas is incubated at each of three

temperatures, 30°C, 37°C and 40°C. These conditions range from fully inducing and indication-capable (30°C, high IPTG) to fully repressing and indication-negative (even induced cells would not turn blue due to the thermosensitive *lacZ* allele) (40°C, no IPTG) Colonies that are blue at any condition are then candidate nuclease genes. The darker the blue color, the greater the DNA-damaging activity.

Individual colonies can then be recovered from master plates that have not been subjected to the damaging condition, to assure recovery of the original sequence, grown in small cultures (10 ml LB with antibiotic) and plasmid preparations made for storage.

Reversing the configuration of expression so that the repressing condition is at 30°C +IPTG and the inducing condition is 40°C - IPTG can be easily accomplished with pLT7K by switching the cloning sites added to the oligonucleotide primers for PCR so that cassettes are in the reverse orientation. This may be desirable to facilitate storage of never-induced colonies. For this purpose strain ER2745 is the preferred strain, since the damage-inducible fusion carries a wild type *lacZ* allele that enables indication at 40°C. In that case, the colonies desired will be darker blue than the normal light blue color.

Further characterization is then carried out on the identified plasmids, either continuing from the replica plate masters or from the archived plasmid DNA following retransformation. Further characterization includes some or all of the following three steps.

Crude extract assay: Clones positive in the DNA-damage screen are grown at in medium-sized cultures (20-200 ml) at 40°C -IPTG (noninducing conditions) in LB + ampicillin to late log phase, and shifted to the inducing condition identified for the clone (usually 30°C + IPTG, but possibly a semi-inducing condition) for four hours. This procedure was successful in allowing expression of an amount of PacI similar to that expressed in the native host, *P. alcaligenes* (D. Byrd, personal communication). Cells are then collected by centrifugation, resuspended in buffer, lysed by lysozyme-EDTA treatment, and clarified by centrifugation.

Crude extracts supernatants are then assayed for nuclease activity in a general screen for 4-6 base cutters, using standard plasmid, phage and viral DNAs such as pUC19, pACYC187, pACYC177, pBR322, M13mp18 replicative form DNA, lambda DNA or T7 DNA at 37- 68 °C. Some 8-base specificities may be detected by this method as well.

DNA digestion patterns are resolved by agarose gel electrophoresis using an agarose concentration suitable for visualization of bands between 200 and 0.05 kb (usually 0.7% agarose and 1.3 % agarose), and detected by ethidium bromide staining.

DNA digestion patterns are then evaluated and the recognition sequence is determined by methods known in the art. Further purification of the endonuclease thus identified may be required for these methods to be applied.

Crude extract supernatants are also assayed in an in vitro screen for enzymes with 8-base sites, using chromsomal DNAs of varying GC-content: *Rhodobacter sphaeroides*, *Escherichia coli* and *Staphylococcus aureus* range from 66% to 34% G+C and are suitable for detecting a variety of enzymes with rare sites. It is usually possible to distinguish between nonspecific nuclease and an 8-base endonuclease, since specific fragments (especially large ones) are not subject to further digestion; even though the fragments are not resolvable on the gel (and the recognition site cannot be deduced), the result is recognizably different from that produced by nonspecific nucleases (which preferentially degrade large fragments). In each case, aliquots of extract are incubated with potential DNA

substrates in the presence of Mg^{++} and resolved on agarose gels followed by ethidium bromide staining.

Isolates that yield a positive result on chromosomal digests but not in digests of standard substrates are then further characterized by searching for alternative substrates, guided by the G+C content of the chromosomal DNA yielding a positive result.

Pulsed-field gel assay: A potentially more-informative assay for 8-base recognition sites relies on separation of total chromosomal fragments on pulsed-field gels. When crude extracts are used for screening procedures, these gels are too cumbersome and too sensitive to other nucleases in the extract to be generally useful.

In standard procedures, the substrate DNA is obtained by first embedding whole cells in agarose plugs. DNA is released from the cells in situ by means of a series of enzymatic treatments and washes that degrade the cell wall. The restriction endonuclease is then incubated with the plug; this usually takes several hours, since the enzyme must permeate the agarose and the remnants of the previous digestions.

In this method the restriction nuclease digestion step consists of inducing expression within the cell, before agarose is added; embedding the cells in agarose and subjecting the cells to electrophoresis on a pulsed-field agarose gel. Controls include: positive control, standard digestion of the host DNA embedded in agarose plugs with purified PacI and NotI; and negative control, samples of the host containing the empty vector, treated in parallel with the experimental samples.

Possible improvements in the strain used for this part of the survey include introduction of a *recD* mutation, which would inactivate the major ATP-dependent double-strand exonuclease of the cell; and introduction of an *xth* mutation that would inactivate the major ATP-independent double-strand exonuclease. A triply nuclease-deficient strain (*endA xth recD*) should be viable but may not stably maintain the plasmid (Niki, *supra* (1990)).

Isolates identified by this method are then carried further, with further purification and overexpression of the cassette-encoded polypeptide, so that conventional pulsed-field analysis can be carried out.

Fingerprinting: Plasmid DNAs prepared from candidate clones obtained by the indirect report assay are fingerprinted by restriction enzyme digestion. Each candidate is digested separately with two to four enzymes with four-base recognition sites: in the preferred example, with HaeIII and MseI to yield a patterns characteristic of the cloned cassette.

Sequencing: All plasmids that result in banding patterns in crude extract or pulsed-field gel assays are then sequenced.

All fingerprinted plasmids are grouped according to fingerprint and two in each class are sequenced. A minimum of three-fold sequence coverage will be required in order to have sufficient confidence to carry out preliminary homology searches.

Sequencing is carried out using the Tn7-based transposition system, GPS™-1 (NEB Catalog No. 1700, New England Biolabs, Inc., Beverly, MA). This system enables introduction of primer-binding sites at random locations in plasmids of interest, rapid mapping of the location of the insertion by digestion with rare-cutters that cleave within the transposon, and sequencing of the insertions within the fragment of interest. With these target molecules, about 20% of transposon insertions will be found within the sequence of interest. No more than 6 suitable insertions are needed in most cases, since cassettes are normally smaller than 2 kb. Two sequence runs (500 bp per run) from flanking vector primers and 12 runs from insertions will yield 7000 bp of raw sequence, approximately 3-fold redundancy. This will be sufficient for primary analysis. Further sequencing can be carried out to obtain high-quality sequence of the most interesting fragments.

Alternative sequencing methods may be used, such as primer-walking, nested deletion construction, or alternative transposon-based methods such as Primer Islands (Perkin-Elmer).

Sequence Evaluation: Homology to genes in public databases will help to exclude candidates for new type II RM genes. Many genes that might be recovered during this procedure exhibit conserved amino acid sequence segments: topoisomerases, helicases, nicking enzymes associated with conjugal plasmid transfer, and transposases all can be found annotated in databases, identified by BLAST or other homology search procedures. Genes for type II restriction enzymes, on the other hand, rarely can be identified in this way. When they can be identified by homology, they are almost always isoschizomers of (recognize the same site as) the enzyme in the database (R. Roberts, personal communication). Thus, the target genes (endonucleases recognizing new specificities) can be expected among those not identified by homology search.

2. Cassette identification: methyltransferase gene acquisition.

In one preferred procedure, the desirable function is a methyltransferase gene, which may be selected or screened for by methods known in the art, described above.

A. The methylase selection method

This may be used if an endonuclease with suitable specificity is available. This method will be applicable when something is known or suspected about the specificity of potential methyltransferase enzymes and a suitable endonuclease is available. Such an endonuclease may be a heterologous endonuclease recognizing a subset of the relevant sites.

B. The methyltransferase indicator method

This may be used if the vector employed is compatible with the strains previously described (Piekarowicz, et al., *supra* (1991); Piekarowicz, et al., *supra* (1991); Piekarowicz and Weglenska *supra* (1994)), with the proviso that the *dinD*::*lacZ* indicator allele resident in the strains identified in (Piekarowicz and Weglenska, *supra* (1994)) are unable to indicate at temperatures above 37°C, so

only the presence of blue color at or below that temperature should be evaluated. Other strains derived from these may be constructed to enable use of other vectors such as pLT7K.

C. Degenerate methyltransferase-motif PCR

The method of may be employed alone, or the degenerate methyltranferase-motif primers may be combined with a repeat-specific primer or primers annealing to the flanking repeats in a single orientation, such as those employed in PCR fingerprinting or cassette cloning as described above.

D. Biochemical methods

Other methods for evaluating the presence of methyltransferase genes include detection of enzymatic activity such as evaluation of ^{3}H-SAM incorporation into specific DNA sequences and may be applied to individual clones or pools of clones.

E. Hybridization methods

Hybridization detection methods such as colony lifts may be employed to detect the presence of genes with high levels of DNA homology to available methyltransferase genes or to oligonucleotides designed based on the sequences of those genes.

The present invention is further illustrated by the following Examples. These Examples are provided to aid in the understanding of the invention and are not construed as a limitation thereof.

The references cites above and below are herein incorporated by reference.

EXAMPLE 1

IDENTIFYING REPEAT SEQUENCES AND OBTAINING CASSETTES

This Example outlines the general strategy for identifying a candidate repeated sequence. It also provides a specific repeated sequence family, probes for identification of organisms containing similar repeats and primers for amplification of the gene cassettes.

A) Cloning of portions of a superintegron array.

The organisms expressing PacI and PmeI were isolated by at NEB (Polisson, U.S. Patent No. 5,098,839 (1992); Morgan and Zhou, U.S. Patent No. 5,196,330 (1993)). These restriction enzymes are made by particular isolates of *Pseudomonas alcaligenes* (ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc.; Beverly, MA) and *Pseudomonas mendocina* (ATCC No. 55181) (NEB Deposit No. 698, New England Biolabs, Inc., Beverly, MA) respectively. The genes encoding these enzymes were identified and cloned using seven steps: 1) PacI and PmeI were purified to homogeneity from *Pseudomonas alcaligenes* (ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc.; Beverly, MA) and *Pseudomonas mendocina*. (ATCC No. 55181) (NEB Deposit No. 698, New England Biolabs, Inc., Beverly, MA) by the methods of (Polisson, *supra* (1992); Morgan and Zhou, *supra* (1993)). 2) The N-terminal sequences of these proteins were obtained by standard microsequencing methods. 3) Degenerate oligonucleotides, designed on the basis of these sequences, were used to obtain PCR fragments encoding these N-termini. 4) The DNA sequence specifying these N-termini was determined from the PCR fragments. 5) Unique oligonucleotides designed from these specific sequences were used for inverse PCR, to obtain larger fragments encoding the entire genes. 6) In both cases, suitable enzymatic activities were identified in crude extracts of *E. coli* carrying the relevant genes under the control of the T7 RNA polymerase. 7) Further cloning of adjacent sequence was carried out, and sequence was obtained of 4.07 kb of *Pseudomonas alcaligenes* ((ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc.; Beverly, MA) DNA and 5.37 kb of *Pseudomonas*

mendocina (ATCC No. 55181) (NEB Deposit No. 698, New England Biolabs, Inc., Beverly, MA) DNA.

Examination of these sequences by visual inspection enabled preliminary identification of repetitive sequences common to both gene segments. Further cloning experiments were aimed at obtaining a complete sequence description of the cassette array residing in *Pseudomonas alcaligenes* (ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc., Beverly, MA), resulting in four segments of contiguous sequence as described below. Routine cloning procedures were from (Sambrook *supra* (1989); Maniatis, et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1982); Raleigh, et al., Current Protocols in Molecular Biology John Wiley and Sons, New York, pp. 1.4.1-1.4.7 (1989); Moore, et al., Current Protocols in Molecular Biology, John Wiley and Sons, New York, pp 2.0.1-2.6.12 (1999)).

In the expectation that repetitive arrays might be unstable in *E. coli*, we initially avoided attempting to isolate large fragments containing PAR elements. Further *P. alcaligenes* (ATCC 55044) (NEB Deposit No. 585, New England Biolabs, Inc., Beverly, MA) chromosomal DNA fragments were obtained from HindIII libraries constructed by cloning size-selected HindIII fragments into the HindIII site of pBR322. Chromosomal DNA of *P. alcaligenes* (ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc., Beverly, MA) prepared by the procedure described in the manual of Qiagen (Genomic tip 100/G (Cat 10243) was digested with HindIII to completion. HindIII fragments were isolated by gel fractionation on agarose gels (0.7%) and fragments between 2 kb and 10 kb were isolated using QIAquick Gel extraction kit (Cat # 28704) according to the instructions of the manufacturer and ligated with HindIII-digested dephosphorylated pBR322.

The rationale for this procedure is that *P. alcaligenes* DNA is GC rich while the HindIII site is AT rich (AAGCTT). Therefore few chromosomal DNA fragments are as small (2 kb and 8 kb) as those identified by Southern blot to *pacIR* and PAR-specific probes (see section C1 for this procedure). Plasmid preparations were made from 108 of the colonies obtained following transformation using QIAprep Spin Miniprep Kit Cat #27106. 95 of 108 HindIII

clones (88%) carried inserts. These were digested with AclI (AACGTT), which cuts within the PAR sequence identified by eye but rarely in the GC-rich *P. alcaligenes* chromosome, and clones were identified that carried exceptionally large numbers of AclI sites. 11% of clones with inserts (11 clones) fit this criterion. Further characterization by PAR-specific PCR (see Section C2) and sequence analysis (below) verified that these did indeed contain PAR sequences.

The high frequency of PAR-containing fragments in the absence of any selection except for size presumably reflects a higher density of HindIII sites within the PAR-containing region than in the chromosome as a whole. We estimate that size selection eliminated about 90% of all chromosomal sequences. If the total genome is 6-8 Mb (Rodley, et al., *Mol. Microbiol.*, 17(1):57-67 (1995); Dewar, et al., *Microb. Comp. Genomics* 3(2):105-117 (1998)) and 10% of this is represented in the size fraction chosen (600-800 kb total), then 100 inserts of average size ~8 kb would be required to cover all of this fraction. A library of this size would of course not contain all fragments exactly once and not all fragments in the fraction are 8 kb. Nevertheless, the incidence of PAR-containing fragments in the library is consistent with the estimated size of the putative superintegron ($\geq$60 kb; 10% of 800 kb would be 80 kb).

Additional clones were isolated in subsequent libraries made by digestion with ClaI and cloning into the ClaI site of pBR322. At this stage instability of large fragments did not appear to be a problem, so the DNA was not fractionated but was cloned directly. PAR-positive clones were identified by PAR fingerprinting by the method described in Section C2.

Candidate PAR-containing clones were sequenced with an ABI377 sequencer using dye terminators. Template generation was by a combination method. In a semi-random phase, a Tn7-based transposon (an early version of the NEB GPS™-1 kit, (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. 7100) was used for insertional mutagenesis of clones, and selected insertions were sequenced using universal primers (PrimerN and PrimerS, (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. OS1266 and NEB Catalog No. 1267) designed to sequence from the transposon.. Sequencing was facilitated by limited mapping of insertions, employing rare-cut sites within the transposon. Vector-

insert junctions of primary clones and of a few deletion derivatives were also sequenced using primers annealing to pBR322 (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. 1204 and NEB Catalog No. 1205).

This resulted in four sequence contigs totaling 59.4 kb, containing 74 examples of the repetitive sequence. These sequences are SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, and SEQ ID NO:4.

B) Formulation of a repeated sequence candidate.

The specific repeated sequences that are likely to signal the presence of a cassette array can be identified by similarity to those found in known arrays such as the VCR elements of *Vibrio cholerae*, or by computer-assisted analysis of existing sequence information. These sequences were identified by the following procedure, employing computerized search procedures (both UWGCG SEQED and DNASTAR EDITSEQ programs are suitable): the 5' end of the repeat was found by searching for the sequence TAACWA; the 3' end of the repeats were found by searching for the sequence CGTTRR; and the additional constraint was imposed that the 5' base of the 5' element should be not more than 200 bp from the 3' end of the 3' element. This strategy identified 18 repeated elements in this contiguous stretch of 14.144 kb. For comparison, a similar search employing the motifs suggested by Hall (5) identified 11 elements; 10 of these were congruent with the set identified by the strategy cited here, and one aligned very poorly in the internal regions with the others identified by either strategy.
Fig 3 shows an alignment of a set of such sequences identified in a part of the *P. alcaligenes* (ATCC No. 55044) (New England Biolabs, Inc. Beverly, MA, NEB Catalog No. 585) superintegron sequence SEQ ID NO:1. The elements were aligned using the DNASTAR MEGALIGN program, by the CLUSTAL method. The alignment shows a majority consensus (third line), a 90% consensus, at which 16 of the 18 elements are identical (second line) and an identity consensus, with which all elements agree. Only those positions that disagree with the majority consensus are shown on the alignment. 48% (42/87) of positions in the alignment are identical in 90% of representatives; the most divergent representative (PARf9) still agrees with the majority at more than half of positions (47/87).

An additional method for identifying such a repeat is to use a computerized comparison algorithm such as UWGCG COMPARE and DOTPLOT, or the DNASTAR algorithm ALIGN with the DOTPLOT subprocedure. The output of these programs will identify off-diagonal similar sequences (Fig 4; window of 30, match of 24), which can then be examined more closely using a program feature (in DNASTAR) or by noting the approximate positions of the alignment and following with the UWGCG BESTFIT algorithm on the local subsequences surrounding the diagonal. The DOTPLOT method identified 18 elements also: 16 of these were identified by the strategy cited here while two of those identified by the motif search were not found by DOTPLOT. More sophisticated computerized search procedures based on these methods may also be developed and employed for this purpose.

A complete set of the elements identified by searching for the motifs as described is displayed listed herein (SEQ ID NO:5 through SEQ ID NO:78 Table 1). In these elements, an additional two bp adjacent at the 5' end have been added to each element, since these bp are conserved in the majority of the sequences, as 5' GC 3'. One additional base has been added at the 3' end, since this bp is also conserved as C in the majority of sequences. The length of each element, and its location in the relevant contig, and the name of the contig in which it is found is also entered in this table.

It may be noted that the individual sequences within the set display imperfect internal inverted repetition (Fig 5 shows an example of potential secondary structure). This property was also observed in "59 bp elements" and VCR elements.

It may also be noted that the PAR elements fall into families of more-closely related sequences. Alignments of four of these families are displayed in Fig. 6A-6D. Knowledge of these families will inform the design of specific oligonucleotides for further procedures such as those employed below.

Once a repeat sequence candidate or family has been chosen, either from among known arrays or by analysis of new sequence, oligonucleotide probes and

primers can be designed for use in Southern blot and PCR experiments, described further below. Examples of these are shown aligned with the consensus of 74 PAR elements (majority rule) in Fig. 7A (Oligonucleotides 1-5 (SEQ ID NO:79 through SEQ ID NO:83; see Table 2) for Southern blot) and 7B (Oligonucleotides 6 and 7 (SEQ ID NO:84 and SEQ ID NO:85; see Table 2) for PCR).

C) Identifying candidate prokaryotic populations.

With the information obtained from one or more array sets, it then becomes possible to screen additional isolates for the presence of such arrays by Southern blot procedures or by PCR.

C1) Southern blot to Pseudomonas alcaligenes (ATCC No. 55044) (NEB Deposit No. 585, New England Biolabs, Inc., Beverly, MA)

A Southern blot (Fig. 8) was carried out using a mixture of biotin-labeled oligonucleotides (Oligonucleotides 2-5, SEQ ID NO:80 through SEQ ID NO:83; see Table 2) as a probe for repeat sequences (PAR elements), and chromosomal DNA of *P. alcaligenes* (ATCC 55044) (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. 585) prepared by the procedure of Qiagen (Genomic tip 100/G (Cat 10243). Restriction digests with 8 different restriction enzymes (SphI, PstI, StuI, NdeI, NcoI, EcoRI, ClaI and HindIII) were carried out according to the manufacturer's instructions (New England Biolabs, Inc., Beverly, MA). Products were subjected to electrophoresis for 1 h at 100 mA in 0.7% agarose with Tris Borate buffer (composition 0.09 M Tris-borate, 0.002 M EDTA, 10^{-4} µg/ml ethidium bromide). The Southern procedure was carried out according to instructions in the NEBlot® Phototope® kit (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. 7550) using Immobilon-S (Millipore cat #MBBU IMS02) membrane, hybridization at 68°C for 4 h, with 2 washes with at 23°C followed by 2 washes with 0.1XSSPE, 0.1% SDS at 68 °C for 5 min. Development was with Phototope®-Star detection kit (New England Biolabs, Inc., Beverly, MA, NEB Catalog No. 7020) chemiluminescent detection according to the manufacturer's recommendations. Fig 8 reveals that multiple fragments in each digest hybridized with the probe, confirming that the oligonucleotide recognized a repeated sequence.

The minimum sum of sizes of hybridizing bands ranged from ~20 (PstI) to ~44 (NdeI) kb, suggesting that a large number of cassettes might be present. Some of these bands may represent doublet or triplet co-migrating species, so the maximum size cannot be reliably estimated.

Alternative possible oligonucleotide sequences might be designed based on specific families of PAR elements. A single oligonucleotide such as Oligonucleotide 1 (SEQ ID NO:79; see Table 2) may be used (data not shown), which may be used to prepare a biotin-labeled probe by starting with an unlabeled oligonucleotide, and labeling it by use of a random-priming kit such as NEBlot® Phototope® kit.

Other detailed procedures may be used for detecting the presence of hybridization between the probe oligonucleotide and the DNA preparation. The Southern blot procedure separates DNA fragments by size, transfers these to a membrane support, denatures the DNA, hybridizes the probe, then separates the hybridized product from the nonhybridized probe (in this case oligonucleotides) by washing. Alternative derived methods for detecting the presence of hybridized DNA include use of arrays of DNA preparations, not separated by size, adsorbed a membrane (dot blots or slot blots (Moore, *supra* (1999)) or microtiter plate (Chaplin and Brownstein Current Protocols in Molecular Biology John Wiley and Sons, New York, Vol. 1, pp. 6.9.1-6.9.7 (1999)) or other support, followed by washing away the unhybridized probe. The configuration of label can be reversed (the target DNA preparation is labeled while the test probe is fixed to the membrane or other support).

Alternative possible detection methods include the use of radiolabeled oligonucleotides (labeled with S^{35} or P^{32} or P^{33}), or of alternative chemical detection methods, such as digoxygenin-based (Roche Molecular Biochemicals Cat #12102201) or fluorescein-based (AP Biotech Cat # RPN 3030) label and detection procedures. Alternative methods of DNA preparation could include purification by detergent/protease treatment followed by precipitation or CsCl centrifugation, or by purification from agarose gels (Moore, *supra* (1999)). Other commercially available kits that rely on gel filtration may also be employed (e.g.

those supplied by 5Prime-->3Prime, or Promega Wizard Genomic DNA Purification Kit, Cat#A1120).

C2) PCR fingerprinting of six *Pseudomonas* species.

A second method for detecting cassette arrays in a population is to employ primers annealing to each end of the repeats separating the cassettes in a PCR experiment (Fig 7B and Fig 9). If the repeats are present and close enough to each other for PCR amplification to be effective, DNA bands representing the cassettes will be observed in ethidium-bromide stained agarose gels following electrophoretic separation.

To validate this method, six species of *Pseudomonas* were tested: *P. maltophila* NEB Deposit No. 515 (New England Biolabs, Inc., Beverly, MA) (PmlI), *P. fluorescens* NEB Deposit No. 375 (New England Biolabs, Inc., Beverly, MA) (PflMI), *P. putida* NEB Deposit No. 372 (New England Biolabs, Inc., Beverly, MA) (PpuMI), *P. lemoignei* NEB Deposit No. 418 (New England Biolabs, Inc., Beverly, MA) (PleI), *P. mendocina* (ATCC No. 55181) (New England Biolabs, Inc., Beverly, MA, NEB Deposit No. 698), (PmeI) and *P. alcaligenes* (ATCC No. 55044) (New England Biolabs, Inc., Beverly, MA, NEB Deposit No. 585) (PacI). Chromosomal DNA made as above (part A) was used in PCR reactions primed by Oligonucleotides 6 and 7 (Fig. 7; SEQ ID NO:84 and SEQ ID NO:85; see Table 2). PCR reactions included 100 ng DNA, 0.2 μmol each oligonucleotide, 1 units of Vent® Exo^+ polymerase, 1X NEB Thermopol buffer in a reaction volume of 50 μl. Thermal cycling parameters were 15 sec denaturation at 95°C, 1 min annealing at 55°C, 1 min extension time at 72°C. 25 cycles were carried out. Products were subjected to electrophoresis for 1 h at 100 mA in 0.7 % agarose with 10^{-4} μg/ml ethidium bromide.

Figure 8 reveals that two of the six species yielded multiple amplification products from this procedure. This confirms the presence of the repeat segments in the correct orientation and at the correct spacing for amplification to occur. It is not possible to assess the number of potential cassettes from this procedure, since some cassettes may be too long to amplify efficiently, especially in the presence of

shorter cassettes that would be amplified preferentially. In addition, some amplification products may represent amplification across two cassettes. In this case, the repeat separating them might be more distantly related to the primers than those at the ends of the amplicon.

Use of a variety of extension times will facilitate acquisition of a maximum variety of cassette products. Multiple reactions employing alternative primer sets annealing at high efficiency to alternative families of repeats will also increase the total yield of cassettes. Primers 8-11 (SEQ ID NO:86 through SEQ ID NO:89; see Table 2) are candidate primers for the forward direction, while primers 12 and 13 (SEQ ID NO:90 and SEQ ID NO:91; see Table 2) are candidate primers for the reverse direction as displayed in Fig. 8

Alternative methods of visualization include chemiluminescent detection of affinity-labeled oligonucleotide primers, fluorescent detection of fluorescently labeled nucleotides or oligonucleotide primers incorporated during PCR, or autoradiography when using radiolabeled oligonucleotide primers or radiolabeled dNTP.

C3) PCR fingerprinting of mixed populations

In principle, it should be possible to apply the PCR-fingerprinting strategy to mixed populations to identify the presence of cassette arrays in a minority of the population. At least two kinds of applications to mixed populations can be tried: PCR using combinatorial pools of individual strains, and PCR using environmental DNA.

C3a) PCR on combinatorial pools:

Combinatorial pools can be achieved by arraying individual strains in addressable arrays, for example, 96-well plates. Pools can be made combining the individual strains, e.g. all strains in one row in one pool; or all strains in one column in one pool; or all strains in one 2D address from a series of plates. Many such pooling procedures have been worked out and will be familiar to one skilled in the art (e.g. (Chaplin and Brownstein, *supra* (1999); Green, et al., Cloning

Systems, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol. 3, pp. 297-548 (1999)).

DNA can be made from these strains individually and the DNA samples then pooled; or the strain cultures can be pooled and DNA made form the pool. Each procedure has disadvantages; in the first instance, a larger number of DNA preps must be made; but in the second procedure, different strains may be differentially subject to cell breakage and DNA extraction, and therefore DNA from some strains will be under-represented relative to others.

In such a pooling procedure, some simple controls will allow assessment of the effectiveness of the overall procedure. For example, a positive control--a strain known to contain an array (such as *P. alcaligenes* (ATCC 55044) (NEB Deposit No. 585, New England Biolabs, Inc., Beverly, MA)--can be included in one pool as a single member while the other members are drawn from negative controls--strains known not to contain a responsive array (such as *P. lemoignei* (NEB Deposit No. 418, New England Biolabs, Inc., Beverly, MA). In another, the positive control can be included in duplicate, in another in triplicate, with corresponding reduction in the representation of the negative control. This will enable assessment of the sensitivity of the overall procedure.

C3b) PCR on environmental samples:

A DNA source of great interest is likely to be DNA isolated from environmental samples (e.g. soil, water, filtered air etc) without first obtaining organisms in pure culture. In this case, PCR from cassette arrays may be even more desirable as a mechanism for obtaining genes in intact form. In this case, the same kinds of positive and negative controls as those described in C1 may be included. In addition to a dilution series of the positive control in a known negative control, other controls should be included. The original environmental sample from which DNA is to be isolated can be divided and a portion doped with a small amount of the positive control strain. DNA extraction from the sample will then include some of the positive control, enabling that portion of the sample to be used as a control for the efficiency of DNA extraction and recovery of known cassettes from a known source. Inclusion of a dilution series of purified positive

control DNA in the environmental sample DNA will serve as a control for inhibitory materials in the environmental sample.

An additional series of controls can estimate the fraction of the sample that derived from eukaryotic organisms. PCR controls can test for the presence of mitochondria, chloroplasts, and nuclear ribosomal DNA genes by methods known to those skilled in the art (von Wintzingerode, et al., *FEMS Microbiol. Rev.* 21(3):213-229 (1997); Sekiguchi, et al., *Microbiology*, 144 (Pt. 9), 2655-2665 (1998)).

D) Cloning the DNA fragments.

Once DNA fragments flanked by repeat segments have been obtained, these can be cloned by standard methods. PCR products can be purified using the QIAquick PCR purification kit (Qiagen Cat No. 28104) or other similar kits. Fragments can be digested to provide ligatable ends compatible with appropriately-digested plasmid or bacteriophage vectors. In the present Example, XhoI and XbaI sites added to the 5' ends of the oligonucleotide primers used for PCR provides directional cloning into pLT7K (Example 2 below) such that a defined orientation is obtained relative to vector-borne expression signals. Accordingly, the use of regulatory signals residing in the vector is feasible. If regulation of expression is not a concern, any vector can be used to clone such cassettes, provided that suitable cloning sites are included at the 5' ends of oligonucleotides used for PCR. Such vectors may be high-copy (e.g. pUC19), intermediate-copy (e.g. pACYC184 or pBR322), or low-copy (e.g. pBeloBAC11) plasmid replicons, or may be bacteriophage replicons (e.g. λgt11). Such vectors may contain expression signals suitable for regulated expression in *E. coli* (e.g. pLT7K; see Example 2), or may be designed for expression in an organism suitable for further experimental test of a particular cassette (e.g. *Bacillus subtilis*, *Streptomyces coelicolor*, *Agrobacterium tumefaciens* or other prokaryotic organism).

The ligated fragment pool will normally be recovered as a clone library of fragments consisting of colonies of the recipient organism containing one or more

selectable marker of the vector on solid media following transformation by chemical methods or by electroporation (Hanahan, et al., *Methods in Enzymol.*, 204:63-113 (1991)).

E) Assay for presence of desired cassettes

The cassettes obtained will encode many different sorts of genes. In many cases, genes encoding functions of one particular kind but with differing specificities have related polypeptide sequences. A particular example of this kind of relationship is the set of genes that encode DNA methyltransferases, which carry out the same reaction (adding a methyl group to a specific base in a specific sequence) but with differing specificities (different particular bases within different particular sequences are modified). These can be tentatively identified by PCR employing primers that anneal to conserved polypeptide motif (Morgan, *supra* (1996)). Briefly, individual colonies or pools of colonies from step D) can be subjected to degenerate PCR by procedures detailed in Morgan, 1996, with modification. Most suitable would be a design in which degenerate primers annealing to the methyltransferase motifs form one end of the amplicon and the other end of the amplicon is formed by one or more of the primers annealing to the flanking repeats. If a PCR product of suitable size is obtained, the relevant colony is likely to contain a gene for a methyltransferase. Plasmid or phage clones from candidate colonies identified in this way can then be sequenced in part or in whole.

Alternatively, plasmid or phage clones from colonies picked at random can be sequenced. Clones with potential methyltransferase genes can be identified by evaluation using DNA comparison algorithms such as BLAST or FASTA, or by means of programs specifically directed to evaluating such similarities (Posfai, et al., *Compt. Appl. Biosci.* 10(5):537-544 (1994)).

Functional tests for specific activities can also be use, as in Example 2.

EXAMPLE 2

FINDING RESTRICTION ENZYME CASSETTES BY FUNCTIONAL REPORT FOLLOWED BY CHARACTERIZATION

The present procedure will allow isolation in expression-ready form of a large number of cassettes specifying a variety of genes with diversity-selected functions. Accordingly, identification of specific clones expressing functions of the desired type is a critical part of the procedure. This example illustrates one way to identify a particular desired function, a DNA damaging agent, and to refine the functional identification until a site-specific doublestranded DNA endonuclease (a restriction enzyme) has been characterized. In addition, this example illustrates that the method is useful even when the desired function is toxic to the cell that expresses it. The procedure of this Example is possible specifically because the orientation of the genes is specified in advance, due to the natural orientation of the genes in a cassette array relative to the repeat elements that separate them.

Accordingly, in one embodiment, the vector employed, pLT7K (Fig 10), can be used to regulate the expression of the cloned cassette fragments even when nothing whatever is known about the identity or sequence of the cassettes individually. In this vector, two levels of control are available: expression is inducible and inhibition is repressible. A T7 gene 10 promoter reads into one side of the cloning site; expression from this promoter is repressed by LacI provided by the vector, as is expression of the T7 RNA polymerase itself, which is provided by the host cells used for expression. Further control can be obtained by the use of pLysP, which expresses an inhibitor of T7 RNA polymerase.

To further reduce expression directed by the cloned fragment, and residual leaky expression from the T7 promoter, tandem λ pL promoter reads into the other side of the cloning site, antagonizing expression from pT7. This antagonistic transcription is regulated by λ cI^{857}, a thermosensitive repressor. At 40°C and in the absence of IPTG therefore, essentially no expression was observed; at 30°C, some leaky expression is seen; at 30°C in the presence of IPTG, moderate levels of expression can be achieved.

The strategy employed in the present Example, an indirect report of DNA damage is used to identify those cloned cassettes that lead to DNA damage, a procedure carried out by subjecting a portion of each clone to conditions that induce expression of the cassettes, and examining the color of colonies thus induced. Those that yield a positive signal are then chosen, and the portion of the clone never subjected to the inducing condition is carried to the next step. This ensures that the DNA damage step does not select for inactivation of the gene identified. The positive cassettes identified at this step (a reduced number) can then be examined in more detail. These are then examined by inducing another portion of each clone and examining the induced portion for three indices of site-specific DNA cleavage. Finally, the clones of interest are sequenced.

A. Reporters of DNA damage for use with pLT7LK.

In order to use the DNA damage indicator strategy for identification of DNA damaging cassettes cloned into pLT7LK, a host strain was required with five characteristics: the T7 RNA polymerase should be expressible after induction; the strain should not contain a lambda lysogen (because it would be induced to express phage-encoded killing functions following DNA damage); it should preferably be highly transformable, in order to obtain a large collection of transformants carrying cloned cassettes; it should express the DNA damage indicator *lacZ*, preferably only following DNA damage--ie with a clean background of white colonies in the absence of induction; and it should not express the major nonspecific endonuclease of *Escherichia coli*, Endonuclease I. This last requirement is needed for clear identification of restriction digest banding patterns in agarose gels, resulting from the action of site-specific endonucleases on test DNA substrates.

ER2745 and ER2746 were constructed by standard P1vir transduction. These strains provide alternative host backgrounds with differing advantages, both useful for the present goal of identifying cassette clones in pLT7K that cause damage to DNA when expressed.

A sample of the ER2745: (F⁻ λ⁻ *fhuA2 [lon] [dcm] ompT lacZ::T7 gene1 gal sulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10*--TetS*)2 R(zgb-210::Tn10* --TetS*) endA1) dinD2::MudI1734 (KanR, lacZ*$^{+}$*)* has been deposited with the American Type Culture Collection under the terms and conditions of the Budapest Treaty on _______, 1999 and has received ATCC Patent Deposit No. _________.

A sample of ER2746: (F⁻ λ⁻ *fhuA2 glnV44 e14- rfbD1? relA1? endA1 spoT1? thi-1 Δ(mcrC-mrr)114::IS10 lacZ::T7 gene1 dinD2::MudI1734 (KanR, lacZ(ts))* has been deposited with the American Type Culture Collection under the terms and conditions of the Budapest Treaty on ________, 1999 and has received ATCC Patent Deposit No. _________.

ER2745 was constructed in one step from an existing strain. The existing strain, ER2566, was deficient in all known endogenous restriction systems (enabling efficient cloning), did not express β-galactosidase, and expressed T7 RNA polymerase under *lacI* control from a chromosomal location (not an inducible prophage). It also lacked Endonuclease I, the major nonspecific nuclease of *E. coli*, and so would be useful for visualizing restriction enzyme activities in crude extracts. The *dinD* indicator was introduced into this strain by P1 transduction from strain ER1992 of Fomenkov, *supra* (1995)), to form ER2745.

ER2746 was constructed in three steps from an existing strain. The existing strain, ER2418, had the desirable property of relatively high induced competence, a property shared by many lined derived from *E. coli* K12 but not present in lines derived from *E. coli*B like ER2745. The allele for expression of T7 RNA polymerase was introduced in two transductional steps: ER2418 x (P1(ER2556) --> TetR (Pro- KanR) to form ER2740; then ER2740 x P1(ER2553) --> Pro+ (KanS TetS Lac- T7RNAP+) to form ER2744. Finally, a *dinD* indicator allele was introduced into ER2744 from ER2170.

B. Cloning the cassettes

Cloning of cassettes was carried out by amplification from chromosomal samples. Total genomic DNA of *P. alcaligenes* (ATCC No. 55044) (NEB

Deposit No. 585, New England Biolabs, Inc., Beverly, MA) prepared by the procedure of Qiagen (Genomic tip 100/G (Cat 10243) as above was amplified using 8 combinations of primers 8-13 (SEQ ID NO:86 through SEQ ID NO:91 respectively; see Table 2): 8+12, 9+12, 10+12, 11+12 and 8+13, 9+13, 10+13, 11+13. The various combinations enable efficient amplification from different families of PAR repeat elements, since the central portion within each family of oligonucleotides (8-11 or 12-13) is varied in sequence. Each of the different versions facilitates annealing to different family members.

PCR amplification was by the procedure of Example 1, Section C2. Amplified cassettes were then digested with 20 units XbaI and 1 unit XhoI (New England Biolabs Cat. Nos. 145 and146, Beverly, MA) in 1X NEBuffer 2 for 1 h at 37°C. Digested fragments were ligated overnight at 16°C with doubly-digested, dephosphorylated pLT7K. Dephosphorylation was for 1 h at 37°C with shrimp alkaline phosphatase (Amersham #E70092Y); ligation was with NEB Catalog No. 202 (New England Biolabs, Inc., Beverly, MA). These ligated libraries were introduced into ER2745 and ER2746 by electroporation, followed by plating on LB + ampicillin (100 µg/ml) and incubation overnight at 40°C. At this temperature, antisense expression is derepressed and in the absence of IPTG sense expression is uninduced, yielding expression undetectable by the DNA damage indicator described below (Section C).

C. Screening for functional report.

The clone library thus recovered under conditions that repress expression of the integron cassettes (40°C -IPTG) to assure viability can then be scored for functional report. Replica plating onto Xgal plates and incubation under semi-inducing (30°C) or inducing (30°C +IPTG) conditions will allow identification of colonies that express DNA damaging functions. Some of these will be restriction enzymes. Individual colonies can then be recovered from master plates that have not been subjected to the damaging condition, to assure recovery of the original sequence.

D. Assessment of clone identity

The DNA damage screen can allow identification of RM genes (Fomenkov, *supra* (1995); Fomenkov, *supra* (1994)). However, other sorts of genes will also be obtained; for example, a single-strand specific nuclease was among the genes recovered using the Endo-Blue method (Fomenkov, *supra* (1994)). Three procedures can be used to identify RM genes. In the first, cells are induced to express the cassette-encoded genes, crude extracts are made, these extracts are used to digest standard target DNAs, and enzymatic activity is detected by production of discrete bands on agarose gels. In the second, clones are briefly induced to express the cassette-encoded gene, then the whole cells are subjected to pulse-field gel analysis. Discrete bands will result from digestion of the chromosomal DNA of the clone-bearing cells. In the third approach, sequencing of clones to allow classification by homology searches.

D1) Crude extract assay

Clones positive in the DNA-damage screen will be grown under non-inducing conditions to late log phase, and shifted to the inducing condition for four hours. This procedure was successful in allowing expression of an amount of PacI similar to that expressed in the native host, *P. alcaligenes* (D. Byrd, personal communication). Cells are collected by centrifugation, resuspended in buffer, lysed by lysozyme-EDTA treatment, clarified by centrifugation.

Digests are of three sorts:

1) a PacI-specific digest using a specific substrate designed to give a diagnostic pattern, for the positive control.

2) a general screen for 4-6 base cutters, using standard plasmid, phage and viral DNAs. Some 8-base specificities may be detected by this method as well.

3) a general screen for 8-base cutters. In vitro screens for enzymes with 8-base sites are more difficult because of the rarity of sites. However, it is usually possible to distinguish between nonspecific nuclease and an 8-base endonuclease

using total chromosomal DNA as a substrate for in vitro digestion with crude. This is due to the presence of specific fragments (especially large ones) not subject to further digestion; even though the fragments are not resolvable on the gel (and the recognition site cannot be deduced), the result is recognizably different from that produced by nonspecific nucleases (which preferentially degrade large fragments).

In each case, aliquots of extract are incubated with potential DNA substrates in the presence of Mg^{++}. Products will then be analysed by agarose gel electrophoresis.

D2) Pulsed-field gel assay

A potentially more-informative assay for 8-base recognition sites would rely on separation of total chromosomal fragments on pulsed-field gels. When crude extracts are used for screening procedures, these gels are too cumbersome and too sensitive to other nucleases in the extract to be generally useful. However, in this case we can to adapt the procedure to our purposes

In standard procedures, the substrate DNA is obtained by first embedding whole cells in agarose plugs. DNA is released from the cells in situ by means of a series of enzymatic treatments and washes that degrade the cell wall. The restriction endonuclease is then incubated with the plug; this usually takes several hours, since the enzyme must permeate the agarose and the remnants of the previous digestions.

The reestriction nuclease digestion step can be bypassed by inducing expression within the cell, before agarose is added. By definition, the candidate clones are known to damage DNA in vivo in regulated manner. Accordingly, a banding pattern should be identifiable using the chromsomal DNA of the cells in which expression of the enzyme is induced. PacI will again be used as a test case. NotI will also be used, since the pattern expected for a total chromosomal digest is already well-known.

Critical steps are: quenching endogenous DNA degradation (especially exonuclease activity) at harvest and during the agarose-embedding process; the

length of the induction; and the degree of induction. Controls include: positive control, standard digestion of the host DNA embedded in agarose plugs with purified PacI and NotI; and negative control, samples of the host containing the empty vector, treated in parallel with the experimental samples.

Improvements in the strain used for this part of the survey include introduction of a *recD* mutation, which would inactivate the major ATP-dependent double-strand exonuclease of the cell; and introduction of an *xth* mutation that would inactivate the major ATP-independent double-strand exonuclease. A triply nuclease-deficient strain (*endA xth recD*) should be viable but may not stably maintain the plasmid (Niki, et al., *supra* (1990)).

D3) Sequencing

Genes obtained can be sequenced. To reduce redundant sequencing efforts, restriction digestion and fingerprinting of large numbers of candidates can be carried out. The recovered genes into sets with similar fingerprints, and two of each are sequenced. A minimum of three-fold sequence coverage is usually required in order to have sufficient confidence to carry out preliminary homology searches.

Sequencing can be conducted efficiently using the newly available Tn7-based transposition system, GPS™-1 (New England Biolabs Catalog No. 1700, New England Biolabs, Inc., Beverly, MA). This system enables introduction of primer-binding sites at random locations in plasmids of interest, rapid mapping of the location of the insertion by digestion with rare-cutters that cleave within the transposon, and sequencing of the insertions within the fragment of interest. With these target molecules, About 20% of transposon insertions will be found within the sequence of interest. No more than 6 suitable insertions are needed in most cases, since cassettes are normally smaller than 2 kb. Two sequence runs (500 bp per run) from flanking vector primers and 12 runs from insertions will yield 7000 bp of raw sequence, approximately 3-fold redundancy. This is be sufficient for primary analysis. Further sequencing can be carried out to obtain high-quality sequence of the most interesting fragments. Other sequencing strategies are also possible.

Homology to genes in public databases can help to exclude candidates for new type II RM genes. Many genes that might be recovered during this procedure exhibit conserved amino acid sequence segments: topoisomerases, helicases, nicking enzymes associated with conjugal plasmid transfer, and transposases all can be found annotated in databases, identified by BLAST or other homology search procedures. Genes for type II restriction enzymes, on the other hand, rarely can be identified in this way. When they can be identified by homology, they are almost always isoschizomers of (recognize the same site as) the enzyme in the database (R. Roberts, personal communication). Thus, the target genes (endonucleases recognizing new specificities) can be expected among those not identified by homology search.

These target genes, for type II endonucleases of unknown specificity, normally can best be identified by adjacency to genes encoding protective modification methyltransferases (R. Roberts and J. Posfai, personal communication). Methyltransferases are recognizable by bioinformatic methods, since conserved motif elements are always present (see above). However, two enzymes that should be recoverable by the present method, PacI and PmeI, are not adjacent to genes similar to any modification methyltransferase, and indeed so far no protective methyltransferases have been identified in the original hosts. Since these enzymes recognize AT-rich 8-base sites and the host organisms contain GC-rich genomes, host protection may be achieved by means of absence of sites.

Accordingly, candidate type II endonuclease genes of special interest will be solo ORFS with no database hits. Candidates adjacent to identifiable methyltranferase genes will be also retained, as will potential isoschizomers, which could have other desirable properties such as those affecting stability.

EXAMPLE 3

GENERAL PROCEDURE FOR EMPLOYMENT OF THE METHOD

Repeats to be sought include those in the public literature (Hall and Stokes, *Genetica* 90:115-132 (1993); Hall and Collis, *Mol Microbiol* 15:593-600 (1995);

Levesque, et al.,*Gene* 142:49-54 (1994); Recchia and Hall, *Mol Microbiol* 15:179-187 (1995); Mazel, et al., *Science* 280:605-608 (1998); Barker, et al., *J Bacteriol* 176:5450-5458 (1994); Clark, et al., *Mol Microbiol* 26:, 1137-1138 (1997); Ogawa and Takeda, *Microbiol Immunol* 37:607-616 (1993); Hall, et al. *Mol Microbiol* 5:1941-1959 (1991); Levesque, et al., *Antimicrob Agents Chemother* 39:185-191 (1995); Sallen, et al., *Microb Drug Resist* 1:195-202 (1995); Sandvang, et al., *FEMS Microbiol Lett* 160:37-41 (1998); Senda, et al., *J Clin Microbiol* 34:2909-2913 (1996); Tosini, et al., *Antimicrob Agents Chemother* 42:3053-3058 (1998)) those disclosed herein (SEQ ID NO:5 through SEQ ID NO:74), and those identified in the genome sequence of one or more model organism of interest. The set of repeat sequences identified in the organism of interest are determined by the method of Example 1. These segments are then made into a multiple alignment, for example using the program MEGALIGN (DNASTAR, Madison Wisconsin) and preferably the CLUSTAL method of alignment within it. Segments thus identified can be grouped into families, for example by means of the Phylogeny facility in the MEGALIGN program, and bushy groups, in which there are many interior branches, are chosen as repeat families. These additional families should direct the design of oligonucleotides for use as probes or primers during application of the method.

2) Identification of a variable class of functions

A function of interest is identified in a taxon related to the model organism of interest. This can be for example ability to adhere to a particular tissue, for example red blood cells or the root hairs of plants.

A relatively large (>50 members) and diverse collection of isolates within the taxon of interest are collected. The diversity of these isolates is characterized by isolation from locations spanning the extremes of the organism's distribution; these extremes may include spatial (geographic) distribution, thermal tolerance, salt tolerance, pH tolerance, O_2 partial pressure tolerance or requirement or host organism identity.

The members of this collection are screened for the presence of the function of interest and its specificity. In this example, it may be done by testing for

hemagglutination ability, with red blood cells of sheep, cows, rabbits, pigs, goats, frogs, and humans as examples of different specific targets, or may be tested with one type of red cell in the presence of different mono- or disaccharides, or following various treatments that alter the nature of the red cell surface. The function is identified as variable in the way that is expected of cassette-encoded functions if one or both of two conditions obtains. First, a large fraction (>10%) is different from the rest, in whether the function is present or absent. For example, 5 or more members of the collection express hemagglutination of the red cells, and the rest don't; or vice versa. Second, the specificity of the function varies: for example, some agglutinate sheep red cells, others goat red cells. This criterion is best satisfied if the number of specificities identified is large, for example >4 different specificities in a collection of 50 isolates.

Variable functions can also be identified by immunological procedures, for example ELISA assays employing sera from animal or human populations of interest, or monoclonal antibodies recognizing variable epitopes in a compound of interest (e.g. a polypeptide); or by cytotoxicity assays, for example employing tissues of different physical or phylogenetic origins; or assays testing inhibition or stimulation of cellular processes such a DNA synthesis or cAMP hydrolysis directly or indirectly, in a context of tissue- or organism-specific effects; or tests of growth on or transformation of varied potential sources of carbon, nitrogen, or energy; or tests of growth in the presence of or inhibition of varied antimicrobial compounds.

3) DNA preparation and determination of suitability for use of the method

A preliminary test of the suitability of the method may be carried out by colony PCR, by inoculating a series of small samples of culture medium (for example in microtiter well plates) with portions of isolates of the taxon to be examined (reserving another portion for storage), growing them, boiling them, and carrying out PCR as in Example 1, Part C2. Other primers designed based on these or other repeat families identified from the literature or in step 1 can also be used. Positive isolates identified at this step by the appearance of one or more PCR product are then carried to the next step.

4) Cassette isolation

DNA preparations from positive isolates is subjected to PCR on a larger scale, employing primer pairs with suitable restriction enzyme cloning sites at the ends as in Example 2: SEQ ID NO:86 with SEQ ID NO:90; SEQ ID NO:86 with SEQ ID NO:91; SEQ ID NO:87 with SEQ ID NO:90; SEQ ID NO:87 with SEQ ID NO:91; SEQ ID NO:88 with SEQ ID NO:90; SEQ ID NO:88 with SEQ ID NO:91; SEQ ID NO:89 with SEQ ID NO:90; SEQ ID NO:89 with SEQ ID NO:91 (see Table 2). Additional primer pairs designed based on additional repeat families may also be designed. Amplification conditions may be adjusted depending on the pairs used.

5) Cassette cloning

The PCR fragments are digested with XhoI and XbaI if the primers of Example 2 and pLT7K arc used; other primers can be used including primers suitable for use with a derivative of pLT7K or similar plasmid carrying other restriction sites at the cloning site.

6) Strain choice

A strain suitable for recovery of cassettes will be one not expressing the function of interest, but in which its presence can be sought. For example, hemagglutinin genes should be expressed in a strain not itself expressinga hemagglutinin that would interfere with the survey. LE392 is an example of an E. coli strain that does not express hemagglutinin activity. For use with pLT7K, the T7 gene1 construct would need to be introduced into LE392; or alternatively, strains such as ER2645, ER2746, ER2566 or ER2744 could be used if they were shown to lack hemagglutinin activity. The strain may be customized to facilitate expression or report of functionality, for example by expressing a protein export system capable of exporting a class of hemagglutinins sought (eg. fimbriae).

7) Cassette identification

In the case of hemagglutination, a functional assay is available, so colonies or pools of colonies can be tested for hemagglutination in microtiter wells, following induction of expression as in Example 2.

Another method of identification would be to design degenerate primers specific for motifs found in particular classes of expected proteins, for example fimbriae, pili, or outer membrane proteins, and use them to perform PCR on colonies or pools of colonies either alone or in combination with PCR primers specific for the flanking repeats, as described in example 2.

A list of motifs characteristic of classes of proteins can be found in the public databases described in (M. Patterson and M. Handel, "Trends Guide to Bioinformatics" Elsevier Science, Cambridge, UK, (1998)).

8) Functional characterization

Colonies specifically exhibiting properties expected of desired gene cassettes would then be characterized by methods appropriate to the particular function identified, for example, in a hemagglutination test by competition with small molecules such as various sugars; by its sensitivity to various treatments such as iodination, heating, freezing, treating with acid, alkali, or alkylating agents or with proteases or nucleases;and by obtaining the sequences of the genes and determining the properties of cells with genes carrying mutations of various sorts including fusions to other reporter molecules such as alkaline phosphatase, beta galactosidase, green flourescent protein or various epitope tags, or obtaining purified preparations of encoded proteins by standard purification methods or by affinity purification by means of polypeptide tags.

WHAT IS CLAIMED IS:

1. A method for the cloning of intact, diversity-selected genes from within gene cassettes, said method comprising the steps of:

(a) identifying repeat DNA sequences which flank gene cassettes;

(b) hybridizing oligonucleotides to said repeated sequences which flank said gene cassettes and amplifying said sequences to provide DNA fragments which contain genes from within the cassettes.

(c) ligating said DNA fragments into a vector; and

(d) transforming said vector into an appropriate strain.

2. The method of claim 1 wherein said diversity-selected genes are selected from the group consisting of:

cell surface antigens such as polysaccharide antigens or polypeptide antigens or secreted molecules; adhesins such as fimbrial proteins, pilus proteins or outer membrane proteins; transporters of small molecules, especially those with narrow specificity; toxins, hemolysins, hemagglutins, kinases and signaling molecules;

detoxifying enzymes such as drug resistance determinants; catabolic enzymes specific for compounds episodically available, excluding those required for central metabolic pathways such as the tricarboxylic acid cycle; enzymes for biosysnthesis of rare sugars, excluding those required in all cells, such as ribose, deoxyribose, and sugars of the cell wall, especially of those sugars that form part of the pericellular envelope.

3. The method of claim 2 wherein said diversity-selected genes comprise restriction endonuclease genes.

4. The method of claim 2 wherein said diversity-selected genes comprise methyltransferase genes.

5. The method of claim 1 wherein said oligonucleotides contain recognition sites which permit directional cloning.

6. The method of claim 5 wherein the DNA fragments are ligated into said vector in an orientation that enables expression.

7. A method for identifying the presence of gene cassette arrays from within a target DNA preparation, said method comprising the steps of:

(a) hybridizing at least one oligonucleotide which hybridizes to one or more of SEQ ID NO:5 through SEQ ID NO:78 to a DNA preparation; and

(b) detecting the presence of a stable DNA-DNA hybrid.

8. The method of claim 7 wherein said detection comprises determining the presence of stable DNA-DNA hybrid by Southern blot or dot blot.

9. The method of claim 7 wherein said detection comprises employing at least two oligonucleotides and hybridizing said oligonucleotides to said DNA preparation, and detecting their ability to support DNA polymerization at the 3' end of the stable DNA-DNA hybrid.

10. The method of claim 7 wherein said oligonucleotides comprise SEQ ID NO:79 through SEQ ID NO:91.

11. The method of claim 7 wherein said oligonucleotides hybridize to one or more of DNA SEQ ID NO:5 through SEQ ID NO:78 or portions thereof.

12. The method of claim 7 wherein the DNA source comprises an individual strain.

13. The method of claim 7 wherein the DNA source comprises a group or pool of strains.

14. The method of claim 7 wherein the DNA source comprises environmental DNA.

15. A composition consisting of isolated DNA primers comprising SEQ ID NO:79 through SEQ ID NO:91 or portions thereof.

16. A composition consisting of DNA primers which hybridize to one or more of DNA SEQ ID NO:5 through SEQ ID NO:78 or portions thereof.

17. A method for identifying gene cassette arrays from a predetermined DNA sequence, said method comprising the steps of:

(a) screening the said predetermined DNA sequence for TAACWA;

(b) screening the said predetermined DNA sequence for CGTTRR;

(c) screening for DNA segments wherein the 5' T of step A is less than about 200 base pairs form the 3' R of step B; and

(d) determining whether the DNA sequence of step C is repeated in the predetermined DNA sequence.

Table 1

SEQ ID NO:5 through SEQ ID NO:39

SEQ ID NO:#	Sequence
SEQ ID NO:5	ATCTAACAAT TGGTTCAAGT CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAGC GTTAGGT
SEQ ID NO:6	ACCTAACATG GCGCTCAACC GCGCTCCCTT CGGTCGCTGG ACGCTGCGCG ATAAAGCCGC GCAGCGCCGG TTAGCTCTAC GTTAGG
SEQ ID NO:7	ACCTAACAAT GCGCTCAACT GCGCTCCCTT CGGTCGCTGG ACGCTGCGCG ATAAAGCCGC GCAGCGCCGG TTAGCTCTAC GTTAGG
SEQ ID NO:8	ACCTAACAAT GCGCTCAACT GCCGCTCACT TCGTTCGCTG GACAGTCAAA AGCTGCGCTT TTGCCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:9	ACCTAACAAT GCGCTCAAAG CGCTCACTTC GTTCGCTGGG ACCGGCGAAG CCGGCCCCTT AGCTTAATCG TTAGGT
SEQ ID NO:10	GCCTAACAAT GCGCTCAAAG CGCTCACTTC GTTCGCTGGG ACCGGCGAAG CCGGCCCCTT AGCTTAATCG TTAGGT
SEQ ID NO:11	ACCTAACAAC TGGTTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:12	GCCTAACAAT TGGCTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCGAA GCCGGCCCCT TAGCCAAACG TTAGGT
SEQ ID NO:13	ACCTAACAAT GCGCTCAACT GCCGCTCACT TCGTTCGCTG GACAGTCAAA AGCTGCGCTT TTGCCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:14	GCCTAACAAG TGGTTCAAAC CGTTCGCTTC GCTCACTGGG ACGGGCTAAA GCCCGCCCCT TAACCAAACG TTAGGC
SEQ ID NO:15	CTCTAACAAT GCGCTCAACT ATCGCTCACT TCGTTCGCTG GACTCGCAAA AGCTGCGCTT TTGCTCGCCC GTTAGCTTAA TCGTTAT
SEQ ID NO:16	TTATAACAAT GCGCTCAAAT CGTTCGCTTC GCTCACTGGG ACGGGCTAAA GCCCGCCCCT TAGCTTAATC GTTAAAT
SEQ ID NO:17	ATTTAACAAT GCGCTCAACT GTCGCTCACT TCGTTCGCTG GACAGCCAAA AGCTGCGCTT TTGTCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:18	CCCTAACAAA TGGTTCAAAG CCGTTCGCTT CGCTCACTCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGAG
SEQ ID NO:19	CTCTAACAAA TGGTTCAAGT CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:20	GCCTAACAAC TCACTCAACC TCGTTCGCTC CGCTCACTGG ACTCGCAAAA GCTACGCTTT TGCTCGCCGG TTAGCTCAAA CGTTAG
SEQ ID NO:21	GCCCAACAAA TGGTTCAAGT CGCTCGCTCC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGG
SEQ ID NO:22	CCCTAACTAG TGGTTCAAGC CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:23	GCCTAACAAA TGGTTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGT
SEQ ID NO:24	ACCTAACAAT TGGTTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:25	GCCTAACAAC TGGTTCAAGT CACTCGCTTC GCTCGTTCGG GACCGGCATA GCCGGCCCCT TAACCAAACG TTAGGT
SEQ ID NO:26	GCCTAACAAT GCGTTCAAGT CGTTCGCTTC ACTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGT
SEQ ID NO:27	ACCTAACAAA CGGTTCAAGT TCGTTCGCTT CGCTCACTCC GGACGCCCGC AAGCTACGCT CGCGGTCGCC CCTTAACCTG TCCGTT
SEQ ID NO:28	GCCTAACAAT ACGCTCAACT ATCGCTCACT TCGTTCGCTG GACGTCCAAA AGCTGCGCTT TTGGCCGCCC GTTAGCTTAA CCGTTAT
SEQ ID NO:29	ACATAACAAT GCGCTCAACT GCCGCTCACT TCGTTCGCTG GACAGCCAAA AGCTACGCTT TTGCCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:30	GCCTAACAAG TCGCTCAACT GCCGCTCACT CCGTTCGCTG GACAGCCAAA AGCTGCGCTT TTGTCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:31	GCCTAACAAT GCGCTCAACT ATCGCTCACT CCGTTCGCTG GACGTCCAAA AGCTGCGCTT TTGGCCGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:32	GCCTAACAAC TGGTTCAAGC CACTCGCTTC GCTCGCTCGG GACCGCGTAC CGCGGCCCCT TAACCAAACG TTGGGC
SEQ ID NO:33	GCCCAACAAA CGGTTCAAGA CCGCTCGCCT TGCTCGCTCG GGACCGGCTA AAACCGGCCC CTTAACCAAA CGTTAGGG
SEQ ID NO:34	GCCTAACAAC TGGTTCAAAT CGCTCGCTCC GCTCGCTGGG ACCGGCGAAG CCGGCCCCTT AACCAAACGT TAGGC
SEQ ID NO:35	GCCTAACAAT GCGCTCAAAG CGCTCACTTC GTTCGCTGGG ACCGGCGAAG CCGGCCCCTT AGCTTAATCG TTAGGT
SEQ ID NO:36	ACCTAACAAT GCGCTCAACT GTCGCTCACT TCGTTCGCTG GACAGTCAAA AGCTGCGCTT TTGCCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:37	CGCTAACAAT TCGCTGCAGG CGCGACGGCC CTGACGGGCC GCGGCCTGAG CTCAAACGTT ATAA
SEQ ID NO:38	TTATAACAAT TGGTTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGC
SEQ ID NO:39	GCCTAACAAA TGGTTCAAGT CGCTCGCTTC GCTCATTCGG GACCGGCTAA CGCCGGCCCC TTAGCTTAAT CGTTAGGC

Table 1 Continued
SEQ ID NO:40 through SEQ ID NO:75

SEQ ID NO:40	GCCTAACAAT AGGTTCAAGT CGCTCGCTTC GCTCACTTGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGT
SEQ ID NO:41	ACCTAACTAG TGGTTCAAGC CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:42	ATATAACAAT TGGTTCAAAC CGTTCGCTGC GCTCACTGGG ACGGGCTAAA GCCCGCCCCT TAACCAAACG TTATGC
SEQ ID NO:43	GCATAACAAT TGGCTCAAGC CGCTCGCTCC GCTCACTCGG ACGTCCGTAA GCTACGCTTC CGGCCGCCCC TTAGCCAAAC GTTAGG
SEQ ID NO:44	CCCTAACAAA TGGTTCAAAG CCGTTCGCTT CGCTCACTCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGC
SEQ ID NO:45	GCCTAACAAT GCGCTCAACT GCCGCTCACT TCGTTCGCTG GACAGTCAAA AGCTGCGCTT TTGCCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:46	CTCTAACAAT GCGCTCAAAT CGCTCACTAC GTTCGCTGGG ACCGGCTAAA GCCGGCCCCT TAGCTTAATC GTTAGAG
SEQ ID NO:47	CTCTAACAAT TGGTTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGG
SEQ ID NO:48	CCCTAACAAA TGGTTCAAGT CACTCGCTTC GCTCGTTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGAG
SEQ ID NO:49	CTCTAACAAC TGGTTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGC
SEQ ID NO:50	ATCTAACAAT TGGCTCAAGT CGTTCGCTTC GCTCACTCGG GACGTCCAAT AGCTGCGCTA TTGGCCGCCC CTTAGCCAAA CGTTAG
SEQ ID NO:51	GCCTAACAAC TGGTTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGG
SEQ ID NO:52	CCCTAACAAA TGGTTCAAGT CGTTCCGCTT CGCTCACTGC GGGACCGGCT AATGCCGGCC CCTTAACCAA ACGTTAGGC
SEQ ID NO:53	GCCTAACACT GCAGTCAACC GGACACCAAA CTGTACGCAG TTTGGTTCCC TCCGCTGCGC TCCGGTGCCG GTTACTTTCA ACGTTA
SEQ ID NO:54	GCCTAACAAT GCGCTCAAAG CGCTCACTTC GTTCGCTGGG ACCGGCGAAG CCGGCCCCTT AGCTTAATCG TTAGAA
SEQ ID NO:55	TTCTAACTAC TGGTTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGCC
SEQ ID NO:56	GGCTAACAAT GCGCTCAACT GTCGCTCACT TCGTTCGCTG GACAGCCAAA AGCTACGCTT TTGTCTGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:57	GCCTAACAAC TGGTTCAAGC CACTCACTTC GCTCGCTCGG GACCGCGTTC CGCGGCCCCT TAACCAAACG TTGGGC
SEQ ID NO:58	GCCCAACAAT GCGCTCAACT GCCGCTCACT TCGTTCGCTG GACGTCCAAA AGCTACGCTT TTGGCCGCCC GTTAGCTTAA TCGTTA
SEQ ID NO:59	GCATAACAAT TGATTCAAGT CGTTCGCTTC GCTCACTGCG GGACCGGCTA AAGCCGGCCC CTTAACCAAA CGTTAGGC
SEQ ID NO:60	GCCTAACTAC TGGTTCAAGT CACTCGCTTC GCTCGTTCGG GACCGCGTTC CGCGGCCCCT TAACCAAACG TTAGGC
SEQ ID NO:61	ATCTAACAAT TGGCTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCGAA GCCGGCCCCT TAGCCAAACG TTATGC
SEQ ID NO:62	GCATAACAAT TGGCTCAAGC CGCTCGCTCC GCTCACTCGG ACGTCCGTAA GCTACGCTTC CGGCCGCCCC TTAGCCAAAC GTTAGG
SEQ ID NO:63	GCCTAACAAA TGGTTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:64	GCCTAACTAT TCAGTCAAGC GGACGCAAAC CCCGCTGCGC GGTCTTTGCG CCGCTTATCT CAAGCGTTAG AT
SEQ ID NO:65	ATCTAACATG TGGTTCAAGC CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGAG
SEQ ID NO:66	CTCTAACAAT TGGTTCAAGC CGCTCGCTTC GCTCGCTCGG GATCGGCGAA GCCGGCACCT TAACCAAACG TTAGAG
SEQ ID NO:67	CTCTAACAAT TGGTTCAGAT CGTTCGCTTC GCTCACTGCG GGACCGGCTG AAGCCGGCCC CTTAACCAAA CGTTAGGC
SEQ ID NO:68	GCCTAACTAC TGGTTCAAGT CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGGC
SEQ ID NO:69	GTCTAACAAT GCGCTCAAAG CGCTCACTTC GTTCGCTGGG ATCGGCTAAA GCCGGCCCCT TAGCTTAATC GTTAGCA
SEQ ID NO:70	TGCTAACAAT GCGCTTAACT GTCGCTCACT TCGTTCGCTG GATAGTCAAA AGCTGCGCTT TTGTCTGTCC GTTAGCTTAA TCGTTA
SEQ ID NO:71	GCCTAACAAC TGGTTCAAAT CGCTCGCTCC GCTCGCTGGG ACCGGCATAG CCGGCCCTTA ACCAAGCGTT AGAT
SEQ ID NO:72	ATCTAACAAT TGGTTAAAAC CGTTCGCTTC GCTCACTGGG ACCGGCTAAA GCCGGCCCCT TAACCAAACG TTAGGT
SEQ ID NO:73	GTTTAACAAC TGGTTCAAGC CGCTCGCTTC GCTCACTCGG GACCGGCTAA ATTCGGCCCC TTAGCAAACG TTAACT
SEQ ID NO:74	ATCTAACAAT TGGTTCAAGT CGCTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAAACCAAG CGTTATGC
SEQ ID NO:75	GTATAACAAT TGGTTCAAGT CACTCGCTTC GCTCGCTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGAT

Table 1 Continued
SEQ ID NO:76 through SEQ ID NO:78

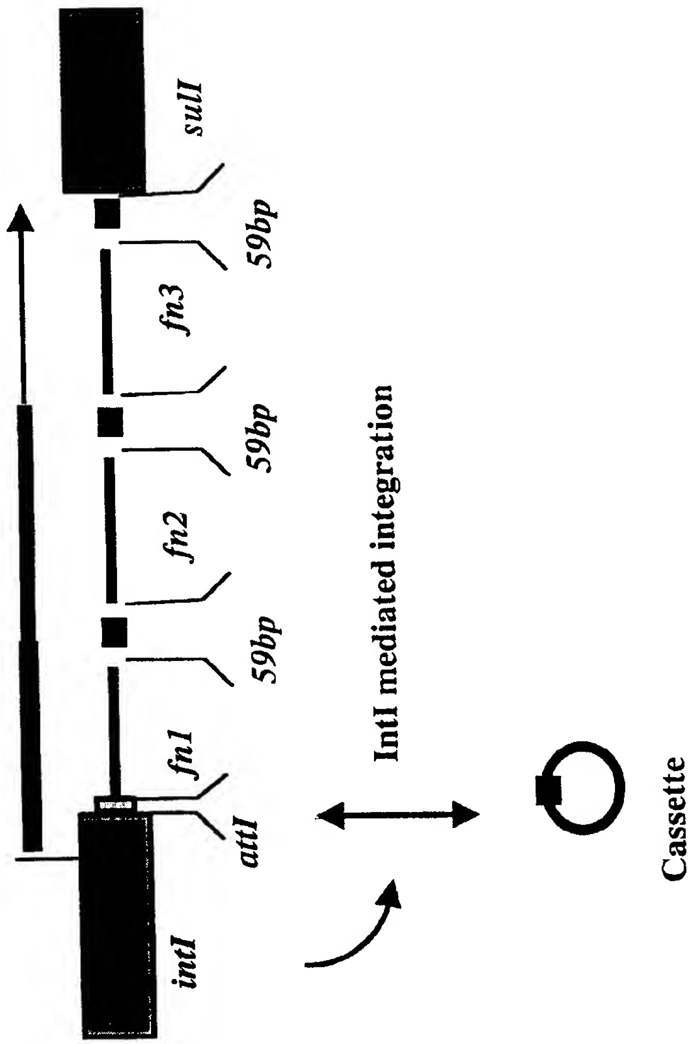
SEQ ID NO:76	GCATAACAAG TCGCTCAAAT CGCTCACTTC GTTTCGCTGG GACGGGCTAA AGCCCGCCCC TTAGCTTATC GTTAGGC
SEQ ID NO:77	GCCTAACAAA TGGTTCAAGC CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGAC
SEQ ID NO:78	CTCTAACAAA TGGTTCAAGC CGTTCGCTTC GCTCACTCGG GACCGGCTAA AGCCGGCCCC TTAACCAAAC GTTAGAG

Table 2

SEQ ID NO:79 through SEQ ID NO:91

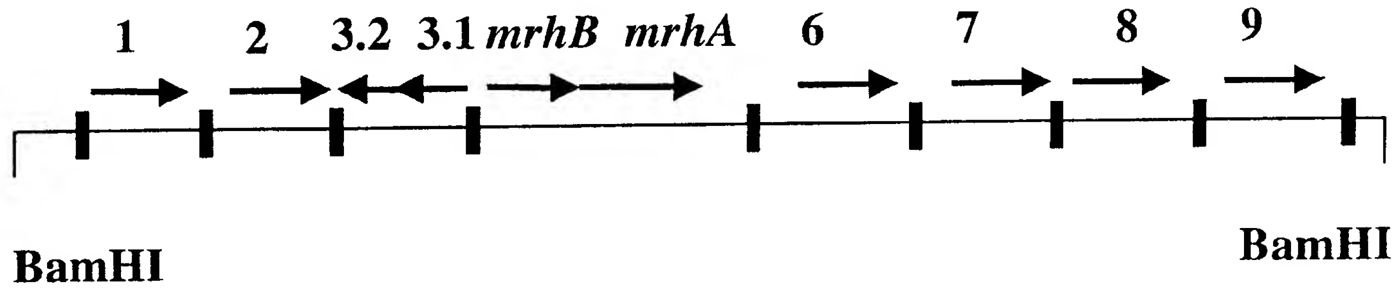
	Name in text	strand	Sequence
SEQ ID NO:79	1	coding	TCGCTTCGCT CACTGCGGGA CCGGCTAAAG CCGGCCCCTT AACCAAACGT TA
SEQ ID NO:80	2	coding	TAACAATTGG TTCAAGTCGT TCGCTTCGCT CACTGCGGGA CCG
SEQ ID NO:81	3	coding	TAACTATTCA GTCAAGCGGA CGCAAACCCC GCTGCGCGGT CTT
SEQ ID NO:82	4	coding	TAACAATGCG CTCAACTGCG CTCACTTCGT TCGCTGGACA GCC
SEQ ID NO:83	5	coding	TAACAAGTCG CTCAACTGCC GCTCACTCGT TCGCTGGACA GCC
SEQ ID NO:84	6	noncoding	GCCCCTTAAC CAAACGTTA
SEQ ID NO:85	7	coding	CCGAGTGAGC GAAGCGAGCG
SEQ ID NO:86	8	noncoding	AAACTCGAGG GTCCCGAGTG AGCGAAGCGA GCG
SEQ ID NO:87	9	noncoding	AAACTCGAGG GTCCCGAGCG AGCGAAGCGA GCG
SEQ ID NO:88	10	noncoding	AAACTCGAGG CTGTCCAGCG AGCGAAGCGA GCG
SEQ ID NO:89	11	noncoding	AAACTCGAGA CCGCGCAGCG GGGTTTGCGT CCG
SEQ ID NO:90	12	coding	TGCTCTAGAC GGCCCCTTAA CCAAACGTTA G
SEQ ID NO:91	13	coding	TGCTCTAGAC GGCCGCCCGT TAGCTTAATC GTTAG

Fig 1 Integron structure



after Hall and Collis 1995

Fig. 2
Vibrio cholerae superintegron fragment carried on pPM147




VCR repeat

Open reading frame (ORF)

ORF2 is similar to Vlp()

ORF3.1 is similar to a plasmid-encoded protein (gi|516610)

ORF3.2 is similar to RelE (gi|42701) and plasmid-encoded proteins

Figure 3A-1
SEQ ID NO:1

ATCGATCAGC CAGACTTTTC GCACACGGGC GGACCTTGGG CGAGTCAGCG
CTATGGTTGG CCGCTGTGGG TTGTCAGTGC CCGTACGCGC AATCTGTTTC
TTTCGCAGGG CATGTCCGGC TGGGCGTTCC GGCCCGTTCT GGTCACCGAC
TCGGCTCTCT ATGAGCGCTA TCTCGCTCTA AGTCAGGAAC TTTGCGCACT
GCTTCGTGAT GCACCGCAGA GCAAGCTCGA AGACCGTGAT TGGTAAGCGG
GGGCTATTCG ATCAGTCTCG GAGCGACCAA ACTCCAGAAA CGACAAGGCC
CTGAAAAAAA AGCAGGGCTT CGTCTTTGCG GGCGAATGGA ATCGGACCTC
TTTCCGCCTC TGCATGTAAC TGGTCTTTGT TTGCCAAATC TGCCTATCTC
ATGCCGGCCA TGTTGGCCAG TGCCTGCATC ATTTGGCCTT TGGTTTCGAC
ACTTTTTCGA CAGCCCTGCT AGACATCCCT CCCTCTGCCC TCGTAACTTC
TGTTCCGATG GTGTCGCTTG GCACTATGGT CTTGTCGAGT GTCGCTTTTC
ATCCAGCCTA ATGCCGCGAT TGCCTCGCTG AGCTGTAGCT GAATCAAGGA
CTTAGCGGAC GACAAGGAAT GTTATGCGAA ACATGTGGCG GAATAAATTA
CGCCGCATGT TTCGTCTACT TATAGTTAGG CTACATATGA GAATCAGCGC
AGACCAGCTT GCTCAAGAAT CACTGACTGA GTTCGGCGTG CTGGCGGCTA
AGCTTCTGGC AACGCGAGAG CTTAGCCAGT TGTCCGAGAA GTTTGGGTAT
GCACTGGCCT TCGGAAGGGA ACCGGCGGCT GCCATAGCTG AGGACCTTGC
TAGGTGCTTG TGCGGACAAA ATGCTTCGCC GGCATCTGAA TACCCCAAAA
TCACCGTTAA GTATTTCAAG GAAAACGAAA GTAGTCTGTT GGCACTCGTA
GAGTGTTATG TACAAATGAC CGCAAGCGCA AACATTCTTT TAGAGCTGGT
TGCCGCACGA AATGGAGAGG CAATAAATCT GTATCTAGAA GGCTTGAGTG
TTGTAGCCTA ACAATGCGCT CAAAGCGCTC ACTTCGTTCG CTGGGACCGG
CGAAGCCGGC CCCTTAGCTT AATCGTTAGA AACCATCATG GATAACTGGT
ACAACACCAT CGAATACCAA ACCCATGTAG CCGAAAAACT AGAGGCACTT
GGAGAAACAA AGTACGACCG CGAGGCTTAT GAATTCGCGC TAGAGGCATA
CCAGTATGCG CCTGAATATC ATGAAAATAT TCCCACGCCG CCTCTCAATC
TTGGGCTCGC GTACCATGTA AGCGCCTTCA ACTTTGCACA CTGCTATGTA
CTTCACGCTA AAGAAGTGTT TGAAGCTCCA AAAGACACAC TGAGCTCCTG
GGGCGTATTT TCCTCAACGG ACATTGGTGA AATTGTTTAT GGTTTAGTCC
GTATTGGCTT GCTGGACCAA GGCCCCGAAG ACAAAAAAGA GCAGTTTGAA
GGGTTGTTTT TAATCACCGA CGTGCTGTGA TGTCTTCTAA CTACTGGTTC
AAGTCGTTCG CTTCGCTCAC TCGGGACCGG CTAAAGCCGG CCCCTTAACC
AAACGTTAGC CACCTCACGA AGATTTGGAG CCCGCGTGAA CAAAGTCGAT
ACAAACAAAA TTAAAACGGA TTTTTCGGCA CGAATTGATG AAAAAAGAGC
GTGGTTTGAT CGTATGGCTA CGCTTATAAG CGGGACAAAC ACCGAGTTAA
CCGACCTTAA TTTTCTTTGC GAGAACTATA TAACATCAAT ATACGTAGAG
CTCGAATGCT TAATATCAGA TTTATTTCAT GGCTACATAA ATAACAACAA
CAAGACCTAC ATGGCGCACA TTCAATCAAA AATCAAGAAC TCCATAACTG
ACAAGTACTC TGCATGGCAC GCCACCCATA CAACATTCGC AGGTCCAGAG
CATATTAATT CAGCACAGCT CAGCACGCTC CTTGATCCAA CAAGCTGGAA
CATCACATTT AAAGACGTTT CCGCAATGAA AGTACGAGCA AAGGAATACC
TTTCCTCAGT ACACGAAAAA AGATTTTCAG GTATATCTGC ATCCGATGGA
GCTCTTATTG ATGCCGCACA TGCAATCAGA AATTGCATTG CACACAACAG
CGAAAGCTCC AGAAAGGTTA TGAACACCAA AATTAAAAGC TTAATTACAG
GCCCAGCTTG CTCAAATGTC GGCCTTGAAC TCACCACAAA TAGTGTGACC
AAAATAGGAA AGTATCTCCG TGCAAATGCT CAGCAAAGCA TGCGAGTGCT
GATTTACTCA GATCGAATAA AATCTATCGG CCTAAGCTTA TAAGTGTGGG
CTAACAATGC GCTCAACTGT CGCTCACTTC GTTCGCTGGA CAGCCAAAAG
CTACGCTTTT GTCTGCCCGT TAGCTTAATC GTTAGGAGGC TCTGCATGAC
TCGTGCAACA GACAGGTTCG AAGAGCTTCT GCAATCACAT GAGTTCTCAG
GGCATATTAT TCGTTGGGTT GCGATATTCG AAGGCCGTCT TGACGGTGTG
TTATCAGTTC ATTTTTCTGG ACTTGAAAGC ACCTATGAAT TCTACGAACT

CATACTTTCC AGGTTGTCTT TCTACGAAAA AATTGAAATC CTGAGAAAAA
TTGATTTTGG TAACAGTCTC AAATCCCAAG AAAATACAGC GCTGCACCTA
GACAAACTGA GGCGATTGCG TAACGCATTG GCGCATGCAG CACACATGCC
ACCTGATGAA ATCATGAAGT TGTGCTCTGA TAAGTGGATA GAGTCCTTTG
TGCTCGGATA TCCAAAGTCC ATTGGCAAAG AGAAAAATGC ACTTGAAAAT
CGGCTATCAC TTCTGTGGAA TTACTGCCAC AGGAGGCATG TAGCAAAAAT
TAAGCAGCTT GCACACGAAC TCAAAAATAC AGAGCAAGCC AACTAATAGA
GTCCAGTTAT ACAGGTCCGT AAATGAGCCG CCTAACAACT GGTTCAAGCC
ACTCACTTCG CTCGCTCGGG ACCGCGTTCC GCGGCCCCTT AACCAAACGT
TGGGCACCCA TAGAAAAATC CTAATGAGAA AACTATTCAT ACCACTAATT
TTCGCCCTGC TATCGGAGAG CTTGATGGCA TCTGAAGCGT ATAAGGACCT
TGAAACACAA GTAACTGAAA AAGCCAGCCT AGCAGTTGCC CAAATGAATG
ACAGAGCAAC TGGAAAGCTC GACTACTCGG AAGAAAGTCT CTATGCAGTA
GAAGAAATGG CAGCGGAAGC AGCTCAATAC AAAGATCAAT TAGATCCAGC
CACTGTAGAC TCGCTTACTC AAGTTCTTGG AAGCTATATT CTTGAGGTTG
CACATAGAAA GCATGGCGGC TCTTACGTTT GGCTTGAATC TGAAAACTCA
CCTGCCTTGG TAGTTGGTGA ACCAGAGTAC AGGCTAGCAC TCTCAACCTT
CGCCAAGGTA CATGGCCGAC TTTCTGGCGA CGAAGCAGAT AATCTTATTT
TCTTCTATCA AGGCTTTTCT GAAAGGCTTA AATCACCATC TCCCGGCATG
AGCGCACTCT ACAAATGAAA CCCGAGTTTG GGGGCCCAAC AATGCGCTCA
ACTGCCGCTC ACTTCGTTCG CTGGACGTCC AAAAGCTACG CTTTTGGCCG
CCCGTTAGCT TAATCGTTAT GCACAATAAA ACATGAAGAC AGCACTCATA
TTTGTAGCTC TAATCTTTCT CTCTGGATGT GACAACTATC AGTCATGCCC
TATAACTGGA AAATGGAAAT CCAACGAAAA GCTAACTTTA GAAAGCATGA
ATGAAACCGG CAGGATAACG GCAAAGCAAA GAGAGATTTT TGAGAACGGC
TTCTTTGGAA AACTAGAATT AGACATAAAT TGCAGTAGCT TCACAACAAT
ACTTGACGGC GTTACCGAAA CCTTTAATTA CGAGATAGTT CGCCAAACAA
AAGATTCCGT CACCGTTAGC TATTACAGCA AAGCGCTGCA AAAACAAGTT
GAGGTCACAT CTATTATCAA CGGAAATTGT TACTCGACAC CTATAGAGCA
GTTAAATTTC AATGAGTATT TCTGCAGAGT CGAGTAGCGC ATAACAATTG
ATTCAAGTCG TTCGCTTCGC TCACTGCGGG ACCGGCTAAA GCCGGCCCCT
TAACCAAACG TTAGGCAAAG GCTCAATGGA TCCCATATTC CATAACATCC
ATAGAAACGA CAAAGAGATT GAGGGCGCTC ATCAACAATG CTCGAGCACA
ATCAATCACT TCATTGAGAT GGTCAAAAAA GGGGGCGAGC CCACCTATAT
GGCAAAGCTA CGTTTTCTTG ACCCTGACAA GTCTGAAAAA GAAGGTAAGA
ATCATATTTT TTATTTGTGG TTATCTGAAG TGCTGTACCA CCCTGCAACA
AATTTACTTT CTGGGGTATT TTTTGAAATC CCTGAAGGCT TTGAAAAGTG
GCACCAAATA GGCCAGCGCC TAGGCTTTGA TCCAGAAGAT GTCTTTGATT
GGATGGTAAT CGACAAAGGT CATGCTAAGG GTGCATACAC ACTAAAGGTA
TCGCGAGAGC GCTTAACCAC CGAGCAAGAA AGAAAAGATT TTGACCGCTA
TATTGGTGTG GCGTCATATG AGTAGCCTAA AATTAAGCGC TCACGCCTCA
GCCTAACTAC TGGTTCAAGT CACTCGCTTC GCTCGTTCGG GACCGCGTTC
CGCGGCCCCT TAACCAAACG TTAGGCGCAA GGGCAATATT GGTCTTCAGC
ACCGAGTCAG GAAACACAAT CACCGAATCA GCGCGGTGTT CCTGAATCGA
ATGGTCGCTG ACAGTTGAGG CCGTTATTTG TGGCCAGCAA AGGAGTTGCT
TTCAGAGAAT GTGCACGTCA CAAATAACTT CCGGGGCCAA AACCGAAACG
CCGTGCGCTC CGCCGGTTAA GCTCGGCGCT GGCATCATTT TCGGCGCTCG
GCTGGGCAAT CTAACAATTG GCTCAAGTCG TTCGCTTCGC TCACTCGGGA
CCGGCGAAGC CGGCCCCTTA GCCAAACGTT ATGCGAGCCA CCATGAATAG
CGAAGAATTA TACAAAAAGG CTATGGAGTT AGAGTCCAAA TGCGAGCATA
AAGCGGCAAT TTCAACTTAC AAAGAAATTG TTAAGAAATC TAACGATCCT
CGACACTTCA TCGCATTCGG AGTTTGCCTC CAAAAATGTG GTCACTGGAA
GCAATCCATC GAGGTATTAG AATCAGGAAT TGCACTGAAG CCTCACTATT
GCGAGGGTGA TGCTCGTCTA TTTTTAGCAA AAGCACTTTT TAAATCAGGC
AAAAAAGGCC TTGCGATAAA GCAATGGCAA CATGTATCAA AAATGCAACC
TGAGTACCCA AGTTATGAGT CTGTGCAAAA TGAAGCCAAG AAAATGCTTG

CACAAAACGC ATAACAATTG GCTCAAGCCG CTCGCTCCGC TCACTCGGAC
GTCCGTAAGC TACGCTTCCG GCCGCCCCTT AGCCAAACGT TAGGCACCAC
ATGCCCTCCA TCAAGTCAGC AAGCCAATAC CAGCGCGCAA CATCGCTCAT
CTTCTTAGTG TCAGGCGCTG CTTGGCTATT CATCGTGCAG TCTTCGTTGC
TGCCATTGAC GGATGTCGCC CGCCAAGAAA TGGTTTGCCT TAATATCGTT
CTTGGTATTG CCTGTTTTGT TATAGGTAGT GCGGCAAAGC GTCAGCGAGA
ATTTCGCTGC CCTGACTGTG GGAACGAAGT AGATCAGAGC TTACCTACAG
AGGGTGATGG CGCCCCACTC CTAAGGCTGT GCAAGCACTG CGATATTCTA
TGGAATGTTG GCAAGACCCC AGACAGTTAA AGTTACCGCC TAACAAATGG
TTCAAGTCGT TCGCTTCGCT CACTCGGGAC CGGCTAAAGC CGGCCCCTTA
ACCAAACGTT AGGCAACAGG GGGTGACATG ACGCAATGTC CAAGGTGCCA
GCGCAATCTC GCAGCTGACG AGTTCTATGC TGGCTCTAGC AAAATGTGCA
AGGGTTGCAT GACTTGGCAA AACCTAAGCT ACAACGCGAA TAAGGAAGGT
CATGCCAACA CCTTCACCAA AGCGACATTT TTGGCGTGGT ACGGCTTATC
AGCACAGCGG CATTGTGGGT ATTGCGGTAT ATCGGAGGCA GGTTTTACAT
CCTTGCACAG GACTAATCCA CGCGGCTACC ACATACAGTG TTTGGGTGTT
GATCGCTCAG ATTCGTTCGA AGGCTATTCA CCTCAAAACG CTCGGCTCGC
CTGTTTTATA TGCAACAGGA TAAAATCAAA CATCTTCAGC GCCAGTGAGA
TGGACGTTCT AGGTGAGGCC ATTTCAAAAG CGTGGCATGG TCGAGGAATT
GCCTAACTAT TCAGTCAAGC GGACGCAAAC CCCGCTGCGC GGTCTTTGCG
CCGCTTATCT CAAGCGTTAG ATGAATAAAA GCCTCCACAC ATAGCCAGCT
TTACGGGAAC GAAGTTGATG CGACGCCTCT TTACTGCCTT AGTATTAATT
TTGATGATTT CTGGGTGCTC CTCCACATCA AAAACTGAAA GCCATAAACA
GCCGCCGAAC AATTCAAGCG ACACGACCGC CATACTGAAA TATATATTTA
CTGTTACCCA TGGCATAGAA GAGGGCTTTG CCAAGTCACT TGAGCAAGGA
AGTTACACAC CAGATGAATA TATGCTCATG CAGCAGGCTT TTAGCAATCT
TGATCTAAAC AGATTAACCA CTCTTTTATC ACCCACTTTA GATAAAAGCA
TGAACACCGC AGACGTAAAA CATTTCATGA TTTTTATAAA ATCTACTGCA
GGTAAGAACT TGCTAACAGC AGGAGAGTCT AGCACTTCTT TTTCCGGCAC
AATGGATAGA GTTCGATCAC TACCCTCCGA ACAGCAGTCG AAAATAAACG
AGTTCTTCCA TGCCAGCTAC ACAAAAAACA CTTTAACAGC CATGGGGGCT
CCAGAGGCGG TACGCATTGT TTATGCATTT GGAGTGGAAT CCATGTGCAA
TTACGCTATG CGCAATAATT TCGAACTGTA TATTTCTATT ATAGAAAAAG
GCAAATGCCA ATAAACCACA AAAAACAAAG GCCGGATCGA TCCATGTGAA
CATCAGCGTT ACATCTAACA TGTGGTTCAA GCCGCTCGCT TCGCTCACTC
GGGACCGGCT AAAGCCGGCC CCTTAACCAA ACGTTAGAGG ATTACATGCC
ATCACTCCAA GAACTCCAAT CGCCGATTGA CTCAGCAATC GTGAACTCCA
TGATTGAGAG CACTCCCGAA ACATGAAGCC AGATTATTCT AACGTTGGTA
CGCGAATCTA ATTCCTCTTG TGTAGGTAAC TTTACACATG AGTTATCCAG
TCCTGAGGGG CATGCACCAG TTGGTCCGGC AGAGAGCTTA TTTGAAGACA
CTTACCAACT CGATGAGTTA TTCTACAGCC ATGGTGAGCG CTTATTCACC
AAGGCAATTT ATCGGGCTAA TGCGGTTGGG GACGGTTGGT CATATCACGC
TGAGTTTGAA TATGCGTAGC ATCCCTCTAA CAATTGGTTC AAGCCGCTCG
CTTCGCTCGC TCGGGATCGG CGAAGCCGGC ACCTTAACCA AACGTTAGAG
ATGGTCATGA ATAAACGCGC ACTAACCTTC GGGCTACTCA TAGCAATTCT
AGCTAGTATC ATTTCACAAG CGCTGCTTCA TGGCCAAAAG GTAATCGCTT
CCGACGTGGC TTTATATACT CCATATTTCC TATCTCCAAT ATTTAGCATG
CTGATACAAG CCACATCGAT GCTGGCATGG GCCATACCTG GACTTTATGT
AGGCTACTTA TGCAAGAACA AGCCAGCACA ACATGGAGCA AAATTGGGGG
CAGCATATGG AATACTTCTT GGATTAATTG TATTCGCAAT GCGAGCTTCG
ACCCAATTAA CGTAGATTCT AAGTTAATAA TCGCAACATC TGCTTTAACA
CAAAAAGCAA AATATTCAGT GCACTTTGCG CTAGTTGCTC CTGCCGGCTA
TCTTCTTGCA AAGCATCGTG CAAATCTCTA ACAATTGGTT CAGATCGTTC
GCTTCGCTCA CTGCGGGACC GGCTGAAGCC GGCCCCTTAA CCAAACGTTA
GGCAACTGAA TGATCACCTG CATTCCGGCA CGTGAATTCC TGCGTAAAGT
ATGCGGCCTG TACGAAGCCT CAACTAATGT AGTTAAGTTG CGTGTATGGG

CTTGTGGATA TGGCATCGCA ATGGATCTAA CTGTCAAAGG TAAATCTGTC
CTTTGTGCGG TTGCGGGAGT ACTCCGCCAA GAGGTCGAAT GCTTTGCTCA
AATTGGGCTT CCGAACGTAA TTCAGTTAGT AGGCGACAAG GCGTCAGAGA
ATCAACTAAA GCTCATAGGC ATGGAACCAC CAATCGAACT TCATATCTCT
CGCGAACAGA GCAGGCTCCA AGTTGTAATC TTGTACGAGG GTCAGGTAAA
GGCTACATAT GTGCTGTCAG CCGCCTAACT ACTGGTTCAA GTCGTTCGCT
TCGCTCACTC GGGACCGGCT AAAGCCGGCC CCTTAACCAA ACGTTAGGCT
TTCAATGAAA ACAGTTCCAG TGAAAATATC AGAAGTCGAA CTAATAGAGA
GTTTTGGGAA ATTCCTGATC AATCAAGACT TAATCGACTA TGAAAATTCC
CACTTCAGTG GCGACGACAA CCATAATGCA GATGTAGCCT TATCTTTAAA
GCCAGGGAAA TGGCCAGGCA TTCAAGTCGA TAAACTACAC ATAGAAGTAA
AGTCACACCA CTCAGAAGAC TCTCAAAACA CCATCAACAA AATATTCGGC
CAATTACTAA AAGAAACCGG AAAGCGAAGC CTCGATAAAG AGAAAGAGTG
CTTAGCTATA TTGTTCCCTT ACGAGCGCGG CGCATGGCCA GGTCGAAACA
ACAAAACAGT AACAAGAATT GAAGGTGAAG CTTATTACCG GAGGGGCTTT
TCGAGAATCG ACAAACAGAC GTTTGTTAAA TTTGGTGACT TGGTCGGTGC
CAAATACATC CTTTCCTTTT CTACAGCATC AAACACATTG AACGTATTTG
AATGGAAAAA TTTCTTAGAT GAGGAATTCA GCCCGATGAT CAGCCTAACA
AATGGTTCAA GCCGTTCGCT TCGCTCACTC GGGACCGGCT AAAGCCGGCC
CCTTAACCAA ACGTTAGACG CACCGGAAAT TTTGCATGGG AAACCAGAAA
ATGGATTTGC AGATAAACGA TACAAAGGTT GAGTGGGTTT CTCCAATACT
GAAGCAATGG ATCAGCATCA ACAAAGAATA CGTCAAGCAA TATGATTTCA
AAGACTGCCT GCACTGGTAT AACGAAAGGG CAAATATAAG TGCTTTTGCT
GGTGCCGTTT GGAAGTCTGG AGGTTTTGCG CTGGAAGAAT ATTCAACTAA
AAAAGGCACC GAAGAAAACA GAGCCAATGG TCGTGTCGAC CTATATTTCT
CCAATGACAA CGAGCAAGCC ATTGTTGAAG CAAAAATGGA ATGGCTCTAC
TTCGGAAAGC GCACAAGACT AGATTTCAAA GAAAAAATAG ATCGTGTAGT
TGAAAAAGCA AAGAATGACA TAATTAACAG CCTGCATGCC AACCCCTACG
ATCTAGGGCT TGGGCTTTCC TTTATTTGCA CATACTGGAA AAAGGGTTAT
GACGCATCCG CCGACATGCA AGCCCTTAGA GCGCTTATGC AAAATTATAA
CTGCGCATTT TATGCAATTT TTGAAAACAG CCCCGACAAC GAAATTGTTA
GCTCAAAGGG CAATATCTGC AACGCTGTGA TTTTAGTTGG GACGGCGCAC
AGCTGAATCG TGTGTGTGCG TCTAACAATG CGCTCAAAGC GCTCACTTCG
TTCGCTGGGA TCGGCTAAAG CCGGCCCCTT AGCTTAATCG TTAGCACTAG
GACTTCCGAC CATCATGAGT GATAGAGACG AATTTTCTGC CCCAACAAAA
AGAGCGCTAG CCGAAAGGAG TGGCTTTAGG TGTTCTTATC TTGGTTGCTC
TAATGCAACC ATAGGGCCTA GTGAAGAATC AGAAACAGCC GTAGCAAGAA
CGGGGGTGGC GTGTCATATA ACTGCCGCAG CGCCCGGCGG AAAAAGGTAT
GACCCAACAT TAAGCCCTAC GGAACGAAGC TCAATCTCGA ATGGTATATG
GATGTGCCAA ACGCATTCAG TTGAAATAGA TAGAGATGAG GCCCGATACA
CATCGACCTT ATTAAATCAC TGGAAAAATA TATCCGAGAG CCGAGCAGAT
TATGCAAAAA ATCATGGCTG GGATATTTTT GACAAATACC CCTTCCTTCA
TATTGACTCG CTAGCCAACA TAGACCTGGC TCTTACCAAA AGCCCTTCCT
CAAATAGCCT TATCGGGAAT GCCATTACAG ACAGCTGCCT CCCTCAACTA
TGGGGTAAAG AGCAATCTGT AATCATCAGA GACCTAATAA TAGAACTTTA
TCGAAATGCC TTCGATCACG GCGAGGCTAG CTCATTCGAA ATATCCATAT
CGGAGCAAAA ACTAGAAATA GTTTACGATG GCAAAAAATT TGACATCTTC
CAACTTCTTG ACCACCAGAA TGCAAACGGT GGCGCCGATA CCTTGCAAGA
AATTGTAGAA AAATATGGCA GTAACTTTGT AGTCAACTAT AGCCACGAAG
GCAACAATAA AATAATAATT CACAGGCTCT CTGACTTTTA CGCGCTTGCA
CCATCCCTCC CGTGCGTAAT ATCACTGAGT GAATACGATG ACAAGGCCCT
AGAGTTAGAC CTGGCTATTT ATGAGCGCTG CGGTGCACTG TACATAATTC
TACCGTTGCA TTTTTGTAGA TCAGATGTCA GGGGGCTAGA GTCGCAGCTA
GCCGCCTTTG AACCTAATGG AAAGCCAGTT TACATTGTAG GCTCAGATGT
GGCAGAGCCT ACAAGAAAAG CAATTATAGA CAGGCTTCCC AACTTCACGT
TCGTCCAAAA GCAATGCTAA CAATGCGCTT AACTGTCGCT CACTTCGTTC

GCTGGATAGT CAAAAGCTGC GCTTTTGTCT GTCCGTTAGC TTAATCGTTA
GGCGCAAGGA GGGACCGTGA CTGAAACTGA GAAAATGGTG GGTAAGTTCG
TCAGCGGTTT TGGCGGGCAG AGATACCGAG AAATTTTTGA AGTCCTCGAA
TCCAGTAACC TTCGCCCACT GGGCAAGTCA AATACTGAAA CATTGCTATT
TCAGCTTCGA GGGGCTGATA GTGAAATGCT AGATATTTTT GCCTTTCGCT
TGGGGCCGCC GCCAGTAATT TCGTTTCCCA AATCATATTG GCTAGGTCGC
CCCAGTGAAT TAAGCGCTCA TCTATCCAAT TTTTCATTCT CGGAAAAGCC
AGCCATAACA GGCCCGGTTT CTGACTCACA GTATTCGGCA GGCCAGGTGG
AAATCACCCG CTCTACTCAT GAGAGGATTA TTGAGGTTTG CAACCGTGTC
TGTGCTTCCC TGCAATAAGC GCCTAACAAC TGGTTCAAAT CGCTCGCTCC
GCTCGCTGGG ACCGGCATAG CCGGCCCCTT AACCAAGCGT TAGATGCAAA
TAACTTGAGG GGCACATGCA AGACTTTGGG TCCAGACGAA ATGCATCATT
AGAGGACAGG GCTGCGGCTG AGTCTGTTAT TGAACGTGTT TATCTTGCGA
TACAGCAGCT TTGCACAGAG ACTGGTGACG TAAGAAATCG GCTTCAAATA
GCCGTTATGA CTCTATTGCC CCTTCAGGCG CGTAACTTCC CCATTGCGTT
GCAGCAAGAC TTCGATTGGA TTGTCAGAGA ATCAACCAAA TACAAATCAC
CATATCCGCA GTTTCGGGGC GACCTTGAAG CAACGATGAT GCGAATAAGG
AACTCAACTG GGCAAAAAAT CGCGCAAAGA ATTTTCAATA TTTACTCGTC
GCTACAAGAC ATTCGAGGTT TTCCCCTGCT TGAATACAGG GCAATAGATG
AGTAAGCATC TAACAATTGG TTAAAACCGT TCGCTTCGCT CACTGGGACC
GGCTAAAGCC GGCCCCTTAA CCAAACGTTA GGTAACCAAG GGAAATTCAC
TTGAGTTGTT ATGTATTGGG CACAAACAAC CGCCATTAAA GGACGGTTTT
ATAGTAAATT TCATCGGACT GTTGAACTAA AATGCTTATA CGCTTTGCTC
TACTACTTGC TGTTATGCTC CTCGCTGCAT GCTCGTCAAA GCAAAATCCA
ACGCCGAAGT GTACTGCCAG CGTCCCCCCG CCCTCTTTAC CCGAAACATC
CACAGTATGC CTAGGGGAAA GATGTAATTG GGAGGTGCTA TTTCCGTCAG
GAAAATACCC TGCATCCACA GAAGGCTGCA GGGCGCCTGT GGTGCAGAAC
CAGCCTTCTT CCTACCCGCG AGAAGCACTT GATCAGTGTA TTGAGGGGTA
CGCTTGGGTA GCGGTTTTTC TGAATGCCGA CGGGGTCCAA ACATCAGCAA
AGGTACTTCA ATCATCGAAT AAAATTTTCG ACAGAAATGC CTTGCTACAG
GCCAGTAATA TATTTTTTGA GCCTATGAAA TGTCAATCCG AGCGTTATGA
TTCCGTTGTT CTGATGCCAT TAAACTACCG CATACTCCCC TAGTAGCGGG
ATTGATCCTT ACAAAATTCA CTACTTACGT CCAAGTTGAA GTAGGCAGTT
TAACAACTGG TTCAAGCCGC TCGCTTCGCT CACTCGGGAC CGGCTAAATT
CGGCCCCTTA GGCAAACGTT AACTATCAGA AGGGCGGTTG ATGTCAAGAT
TTGCGCTCGC GTTGATTCAC GGAGTACCAA CGGGTTTTCT TGTCATTTGT
ACTTTGTTTG TCTGTTTCAT CTACCTCAAC CGATTCGAGA AAGTTGGAGG
ATACTCAGAC GGGTGGGGTT TTGTTGGAAG AGTTGTCTGC GCATCTATAG
CTATGGTTTT CGTGTCCGCA GTTGGCCATC TTCTTATTGA AGCGGCAGTC
AACTGGGGGC TGCAGCAGCT TGGTTATGAG CTGCCAAACT ATGAAAAAAG
AAGGACTTGT AGTAGCTGCA AGCCGAGCAC TCCAGGTGAC TACATGTTCG
GCTTGCTCCT CGGGGGTGTG CTTGGCGCCG GCTCGGCAAT TTGGCTCTGG
ACGCGCCTGG CGCTCCGATA TGCGCTGTTT CGCGGCGAAA ACTGATAGCT
GAACCTTCCA TCGAGGAGAT GCAAAAGCGC TGCTGCGCGC CATCTACAAA
GACCCGAAGC ACCTCATCCA GGCGCTCTCA GCCCGAGCCT GACTGGCTGT
GGCTATCAAC ACCTCTTCGA TACCACTACC CGCCAGAAAC GACAAAGCCC
TGCAAAAAGC AGGGCTTTGT CTTTGGGGAT CTGGAGCGGG CGAAGGGAAT
CGAACCCTCG TCATGAGCTT GGGAAGCTCA GGTAATGCCA TTATACGACG
CCCGCTCGGG CGGCTGACTT TTTACCAGAA TCGCCCGGGA AGGTGAAGCC
GGGCGCGCGT CTTGCGCCCG TTTTATTGCC GGGCGCTTCA TAGCGCCACG
GCCCGTGGCT CTCGTTCCAC GCTGCGTGCG TGGCCCTGCG TGGGTGCCAG
CAGGAAGGCC AGCAGGGCAT CGCGGGTCTG CATCCAGGCG GCCTTGTGTT
CCATGTCGAG GAAGTGGCCG GCCTGGGCGA TGGTGCGGAA CTCGCAGTGG
CGCACGTACT GGGTGAACAG GCGCGCGTCA GCCGGGGTGG TGTACTCGTC
CCACTCGCCG TTGACGAACA GCAGCGGTAT CTCGATCTGC CCGGCGAAGC
TGACGCAGGA GCGCCCGCCG TTGTTCAGCA CGGTTTCCAC GTGGTGACTC

ATTTGCTCAT ATTCATAGCG CTCCAGGCCG GTGACGTGTC GATGGTTGTA
GCGCTTGAAC AGCGAGGGCA GGTGCTTGCC GATGGTGCCG TTGAGCACCA
TGCCGATGCT CTCGCGGTCG CACTCGCGCA TCACCACCAG GCCGGCGCGC
AGGTAGCCGA GCATGGCGCT GTTGACGATC GGCGAGAAGG AGTTGATCAC
CGCACGCTCG ATCCGCGATG GACGCCGGGC CAGCGCCTGG AGGGTGGCGA
TGCCGCCCCA GGAGAAAGGA CAGCACGCTG TTCGCAGGCG AAAATGTTCG
ACCAGCTCCA GGAAGATGTC GGCTTCTTTC CTCGCSGCTG AAG

Figure 3B-1
SEQ ID NO:2

AAGCTTCTGG TACGAACCTG GGGGCGCTCC GGCACGCACA AGGGCATCGA
CATCTTCGCC CGCCAGGGCA CCCCGGTGCT CGCCCCCAGC TACGGCATCG
TGGTGTTTCG CGACGAGCTC GACATGGGCG GCAAGGTACT GCTGATGCTC
GGCCCCAAAT GGCGCCTGCA CTACTTCGCC CACCTCGACA GCTACAGCGC
CCTGCCCGGC CAACCCGTAC TTCCCGGCGC CCCACTCGGC ACGGTAGGCA
GCACCGGCAA CGCCCAGGGC AAGCCGCCCC ATCTGCACTA CTCGATCGTC
ACCCTGTTGC CCTATCCCTG GCGCTGGGAC AACAGCACTC AGGGCTGGAA
GAAAATGTTC TACCTCGACC CCACGCCAAT GCTGAACGAA GCGGCAGTAG
ACAGCCGAAA AACCAGCCAG TAGCGTCGCA GGGGAATGCA CCACCGGTCT
TGCCCGATCC GCCTGTCCTT TTACCAATCG CAGAAGAGTC GCTTTTGTCG
AATCGCCTGT GAGGAAAAAC AAGGACTTGC TGGACGACAA GGAACGTTAT
GCGACACAAG TGGCGGAATA AATTACGCCA TTTGTGTCGT CTACTTATAG
TTATATGCTG ATCTAGATAT GAAGTACAAA AACATAAAAT CAGCAATCCA
CAATTTCGGG CACAGCTTTG TAAGCTCAGT GAACTATGTT GACCATGATT
TCGTTGCCGA CGAAATTGGG AAGATTCACA AGAAAGGCTA TGATATTGAA
ATAAACTGGC TTACAAGGGA GTTCAAGCCC GCTCAGCTTG AGTCAGAGAG
AATAAAAAAA TCAATTGGTT ATTGGGGTGA CAACCTAAAG AAACATTGTG
CATCCCATAG CGTAAATCTG GAAAATCTAT GTTCTTTATC GTTTATCTGG
CCGACAGGTC AAAGTAAATA CATGCATGCC ATTGACGACA AAGGCACAGA
ACACAAAATT TACATCAATG AAGCGCAGTG ATACGCATAT AACAATTGGT
TCAAACCGTT CGCTGCGCTC ACTGGGACGG GCTAAAGCCC GCCCCTTAAC
CAAACGTTAT GCGAGCCACC ATGAATAGCG AAGAATTATA CAAAAAGGCT
ATGGAGTTAG AGTCCAAATG CGAGCATAAA GCGGCAATTT CAACTTACAA
AGAAATTGTT AAGAAATCTA ACGATCCTCG ACACTTCATC GCATTCGGAG
TTTGCCTCCA AAAATGTGGT CACTGGAAGC AATCCATCGA GGTATTAGAA
TCAGGAATTG CACTGAAGCC TCACTATTGC GAGGGTGATG CTCGTCTATT
TTTAGCAAAA GCACTTTTTA AATCAGGCAA AAAAGGCCTT GCGATAAAGC
AATGGCAACA TGTATCAAAA ATGCAACCTG AGTACCCAAG TTATGAGTCT
GTGCAAAATG AAGCCAAGAA AATGCTTGCA CAAAACGCAT AACAATTGGC
TCAAGCCGCT CGCTCCGCTC ACTCGGACGT CCGTAAGCTA CGCTTCCGGC
CGCCCCTTAG CCAAACGTTA GGGGCCAAGA TGGATCTTCG CCAGACAAAG
CCAATACTAG TTACAGTCTT AGCCACTGCC TTGGTGCCAT TGGTTTTTGG
CTGGTATGCG TATTGGGAAA ATCCTCAAGG CATACTTTTG TACACTCCGG
TGGCCGGCCA TCCCCATCCT CAGGGCTCTC CAGCATTTCC TATTGGAGTA
ATGGTTGGGC TGGCCGCTTC ATTTCTGCTC TCTTTGCTTT TTGTAGGCCT
AGGGGGAATC GCTGCATACA TAGCAAGTTC AGTGAGCTCA AAGGCTAGGG
CTAAGCTGTT TTGCAAAATC GCAGTCACAT CCCTGGCTAC TTCAACTATA
GGAGCTGCAG TCTATGCAAT GCTCCCCTAA CAAATGGTTC AAAGCCGTTC
GCTTCGCTCA CTCGGGACCG GCTAAAGCCG GCCCCTTAAC CAAACGTTAG
GCAGCACATA TGACTCGTTC GTGCCTATAC ATGTTTATCG CCTCAGCCTT
GATAGCGTGC GGCGATCCAC CTCTATTGGT TACGCCACTG CCAAATGGCT
ACAATTTCCA TTCCAACGGC GGGGAGTTTG GCTACATCAA GAATCCAGAT
GGATTAAGGC TCGCCGAGTA CTTTGGTATT CGTAATGATG GTCGCGAAAC
CTGGTGCACT GACTTTTCAT GGGAAAGCGA TATCGTCATT TGTAAGCTTA
TTGAATATAG CCAGCATGGA TTTGACGCAT CGCATACAGA GTTTTCTGTA
CTTGACACAA AAACTAGCGA GGTTAGGGTA TTTCCCGATC AAGCGTCTGC
TCAAAATTTC TGGGCCGCAC GCTTTAATTC AGGACTACCT CAGCTTCACC
GGCACTACCC TTCAACCTCA GAGAAGTAAT ATTTTGTGTG TCAGTGCAGC
CTAACAATGC GCTCAACTGC CGCTCACTTC GTTCGCTGGA CAGTCAAAAG
CTGCGCTTTT GCCTGCCCGT TAGCTTAATC GTTAGAGGCT TATTTAGCTC
ATGCGCATAG ACATAGACTT TTCAATATTC ACGCTCGCAC CGTCGACCGA
AGGCGTAATA TCAGGAAAAA TCGAGGTCAG TGAACTACCT AGAACTGGCG

AGATAATTTC ATTCTCCTTT GCGCCAAACA AGTCTAAATT CCCGGCAGAG
CCAAGATTCA ACCCGTTGCT TAAAGTTGAG AGAGTGATTC ATAGCGTAAA
TGGTCAGAGT CCAGCTCTTC AGTTAGAGAA TCTGATGCTA CCAAACAGAG
AAAGTGTCGC TGAAGTCACT GCTTTCCTAG AGCAAGGCTT TGGCCTATTT
TTCAGCCCAA CCGGTGAGTA ATCCTCTAAC AATGCGCTCA AATCGCTCAC
TACGTTCGCT GGGACCGGCT AAAGCCGGCC CCTTAGCTTA ATCGTTAGAG
GTCAGCACAT GGCAGTGCAG CAACTCGGGC CAACCACAGT ATCCGTAACC
GAATTTGCAT GGGACGGAAG CGATCTTGGA AATACTGAGG CCAATGAATT
CTGGTCACAG CTCTCTGCTC AGCTTCAAAA AATAGCTATC TCTGAGTTTT
TAGCTGGCAA TCGCCCCAGC AGCATTCTTC GCAACGACCC ACGAAACATT
ATTGTTCTCT CATTTTCGGC GCCGCCAAAG TTCATTAAAA TCAACCACTG
GCTCTCTGCG TGTACACACA GAATTTCAAC ACGGAAATTA CTGCTACGAC
GGAAACGGCC TGTACTTACG AAAATTTAGA GTCTGGCGAC TTTCTTGCAT
TCGACACAGC GGCGTTGGTG CATGCCCTCT AACAATTGGT TCAAGTCGTT
CGCTTCGCTC ACTGCGGGAC CGGCTAAAGC CGGCCCCTTA ACCAAACGTT
AGGGCACCGC GCATGAGAAA TGAAGACGGA ACCTTTTGCA AAGACTGCCA
CCATCAACTT GATGAAACAC TAGCATCTAG CGCAAATTAC TCATGCCCCA
ACTGCGGCTC CACAAAAAAA TACATGAACA TGTCCATCAC TGATGGAATT
GGCCTATACG ACTCTTTGGG TGCCCAAGCT AAAGATCCAA GTTACCCGGC
AAAAGAAAAA TCAGATGGGA AACATTTGTT GGCTGGGAAC GCAGTCATAA
ACTGCAAAAA ATGGTTTACA AGACAAGAAC TATCGATCGA ACCAATGACG
CATACCAAGA AATAGTAGTC GACCTTAAAA CAGGGGGAAT AATTCATCAC
TGTGAAGAGC CACTTTCAGA GCAYTTKGGC CATGGCACCG CAAAACCAAA
GCCCTAACAA ATGGTTCAAG TCACTCGCTT CGCTCGTTCG GGACCGGCTA
AAGCCGGCCC CTTAACCAAA CGTTAGAGGT TACCTGTGAC AGATTCGCGC
CCGTTACTGA TCCCTGCCTC GCAATATGAT ACGAGCGTTC TTCTCGCCGA
ATGGCAATGG CTCACCCCCA AAACGGATAC GCCACTTTTT ATTTCCATAT
TCGGAGACTG GGTATTTGGC AACCCCAATG GAAGTTTGTG GGTTCTTTCA
CTCCTAAAAG GCACTTACGA GCAAGTAGCC GCAAACTCTA ACGAGTACAA
CACCCTCAAC AAATCGGCGG AGTGGATTGA TCAAACATTC ATCGCCAGTT
GGCAGTCTAT TGCCGCAGGC CATGGGTTAA TCCCAGAACC AAACCAATGC
CTCGGCTGGA AGGTTCACCC ATTATTAGGT GGAAGTTTTG AGCCAGCCAA
TCTCCAACTC TTCAACATGT CGGTGTATCA ATCGCTTATG GGTCAACTTC
ATCGACAGCT TAGCCAAAAA CAAACCCCGG CAAGTAAAAA ACCATGGTTC
CAGTTCTGGT AACCTCTAAC AACTGGTTCA AGTCGTTCGC TTCGCTCACT
GCGGGACCGG CTAAAGCCGG CCCCTTAACC AAACGTTAGG CGCAAGGGCA
ATATTGGTTA TTCAGCACCG AGCCAGGGAA CACAATCACC GCATCAGCGC
AGTGTTCCTG AATCGAATGG TCGCCTGACA GTAGAGGCCG TTATTTGTGG
CCAGCAAAGG AGTTGCTTTC AAAGAATGTA CACGTCACAA ATAACTTCCG
GGGCCAAAAC CGACACGCCG TGCGCACCGT CGGTCAAGCG CAGCGCTGGC
CTCACTTGCA GCGTACGGCT GGGCAATCTA ACAATTGGCT CAAGTCGTTC
GCTTCGCTCA CTCGGGACGT CCAATAGCTG CGCTATTGGC CGCCCCTTAG
CCAAACGTTA GGCCAACATA CTCAACGCAT GAAAACAAAA TATCACATAA
ATATAATTAT ATTTCTCGAA ATCATAATTC CTTTAGCACC AATAATTTGG
GCAATTTTCA CTCAGTCAAG CCCCGGCTTT GGCCCAACCC TTATATCAAT
GCTCATCCTG CACATCGTCG GACGAATAAT TAGCCGAAGC ATCCCTGCCA
GCTGTGACTC ATGTGCTGAA AAAATAAAAC CCAAAGGAAC CTCCGCAATC
TACTACAACT GTCAAAAGTG TGGATTTAAA TACTCAAAAA CACTTAACAG
CAGCAAAAAC TTCCATAACC ACTAACCAGA AAATCACTAA GGCGCCATCA
TGTTATAAGC GCCGTAAGCA CTAAAGACTT GTACAAGCCT AACAACTGGT
TCAAGTCGTT CGCTTCGCTC ACTGCGGGAC CGGCTAAAGC CGGCCCCTTA
ACCAAACGTT AGGGCACTCA ATGCATCGCT TCCTAGCCAC ATGCCTACTA
GCTACATCTA TTAAGGCATA CGCAGAACCT GAAAATAATA TCGACTGCAG
CAACGCATTC TCAACGCCGG ACATTGAACA TTGCGCATCA ATCTCTCTTG
AGAAAACAGA GAAAGAGCTA AATTTAGCAT ATCAAAAATT AGTCAAAGAC
CTTTCTCAGC CAAACAATGA ATACGAAAAT TTCACCGAGT ACAGGAAAAA

ACTTTTAACG GCTCAAAGAG CATGGATCGC GTTCAGGGAA GCAAACTGTG
CCACTCAGTA CGAAATGCAC AGATCTGGCA CTATTCGCAA CAGCATCTAT
CTAGCCTGCA AAGAAAAGCG TGCCAAGCAG CGAATAAAGG GAGCTTCAAA
ATTATGCTCC GTACTAGCCC TAACAAATGG TTCAAGTCGT TCCGCTTCGC
TCACTGCGGG ACCGGCTAAT GCCGGCCCCT TAACCAAACG TTAGGCCGAC
AATCGCAATT CCTAGGACTG CACGTGAACT GGATCCGCAA AATGTTTCGG
CGCACAGCAC TAGCGCCGCC CCAACATCGC GAGGACGAAG CTGTCAGTAC
AAGCCAAGAA GGAACGCCTC CCTTTCGTCA TTTGACAGTT GAGAATTCAT
GGGGAAGTTG AGGGCGGAGC TATTCCTTCA GTCACCACCC GAGAACATCC
TCAGAAGATC TGTTTCTCGT TTGGCGTGCC TAAGTTCGGA TGGTCAACGT
TCGAGATCCA TTTCGTCGGA AATGGCCACT TCATCTGCGG CATCTCTGAC
ACTCCAAATG ACTTCTACGG TGACTTGGCT ATCGCCCTGG CTGAGCAGAA
AAGTTCTTTT TCGGTAGCGG CGCACCTTGA GCCTGAGACC TTTGCCTTCT
ACATCGTTGA TTCGACAATG TACTTGTGCA AGTTCGATGA ATTCGACGAT
TATGAGTCCG CCGCCGAAAG CCACGAACAG TTGGTCTCCC ACAGCTTTAT
GTCCATTGAA GTATCTAGGG AGTACTTTCA GAAGTCTCTC AGGACCTTGG
CCGTCCAATG GCCGGATACG CCTTCAAGAG ACTGGGCGCA CCCATTTCCA
CGTGCGCAGA TTGAAGGCTG ACTGCCTAAC TATTCGCTCA AGCGGGCAGC
GTTAGGCGCC CTCATTCGGA GTCACGCTAT GGCAACCCGA GAAGAAACAG
AAGTAGCCAT TGCTGCTCTT CGCAGCGAAC TCAATGGCAA CGAATCGGAA
TACAGCTTTC ACATTCCCGG TTGGGCGCCA GAAACATCAG TCATGGGATT
TCGCTGGATG CAAAGCCAAC TGTGGGAAGG CTTCTACGTA AGCTATCGCG
TAGAGCACTC GGCCAAGCGC GTCGAATTCA AGTGCTGGGA GTACGGCGAG
CCCGAGCCGT CTTGGCTGCA AGTTGGCTAG GGGGCCGGCA AGATGCAATC
GCGGCGAGCG CCTAACACTG CAGTCAACCG GACACCAAAC TGTACGCAGT
TTGGTTCCCT CCGCTGCGCT CCGGTGCCGG TTACTTTCAA CGTTAGGCAA
CTCAGATGAG TGCTCCAGAC GCAGAACTTC TCGCATTGTT AGCCTACCGA
ATGGAAGCTA TTTCCATTGG GCATTTGGCA TTACGCCATC ACATGACGTG
GGACGAAACA CCTTCAATGG AGGTGTACTT CAATGGCATA CAAGTACTCG
AGGGAAAGGC CACGGGTTTC ACTAATGCAG CCATTGAGTC CGCAATTATT
CATTGCAGGG CAATCCTTGG AGTTTGTTGG GCTGCAGTCC TCCAGACACT
CTTCCACAGA AATTGCAGAG CGCACTCGAC GCAACAATCC CGATGACTAT
GGCATTGAAA GCTTCAATGG CTTATCAATG CTAACCAAGG AAAAAGCACT
AGCCTACTAC TCTGGCGAGC TGCCAGAAGC GGAAGTTGCT CTAGCGCTCA
TATTCCACTC AGCGAACAAA GGGCTTGCAC ACACTACAGT GTCCTTTACG
CGTGACAGTG GCGACGCCCA CCTGATGGAA ATTGCATTTC GCATCGTACC
AATCCTGCTT GTAAATGGCT TCTACGCTCC ACTGGAAATC ACGCCACCAA
AATATGAACT GATTTCACGC CCAAGAGTCG CCATAACAAA TGGTTCAAGT

Figure 3C-1
SEQ ID NO:3

AAGCTTCTGG TACGAACCTG GGGGCGCTCC GGCACGCACA AGGGCATCGA
CATCTTCGCC CGCCAGGGCA CCCCGGTGCT CGCCCCCAGC TACGGCATCG
TGGTGTTTCG CGACGAGCTC GACATGGGCG GCAAGGTACT GCTGATGCTC
GGCCCCAAAT GGCGCCTGCA CTACTTCGCC CACCTCGACA GCTACAGCGC
CCTGCCCGGC CAACCCGTAC TTCCCGGCGC CCCACTCGGC ACGGTAGGCA
GCACCGGCAA CGCCCAGGGC AAGCCGCCCC ATCTGCACTA CTCGATCGTC
ACCCTGTTGC CCTATCCCTG GCGCTGGGAC AACAGCACTC AGGGCTGGAA
GAAAATGTTC TACCTCGACC CCACGCCAAT GCTGAACGAA GCGGCAGTAG
ACAGCCGAAA AACCAGCCAG TAGCGTCGCA GGGGAATGCA CCACCGGTCT
TGCCCGATCC GCCTGTCCTT TTACCAATCG CAGAAGAGTC GCTTTTGTCG
AATCGCCTGT GAGGAAAAAC AAGGACTTGC TGGACGACAA GGAACGTTAT
GCGACACAAG TGGCGGAATA AATTACGCCA TTTGTGTCGT CTACTTATAG
TTATATGCTG ATCTAGATAT GAAGTACAAA AACATAAAAT CAGCAATCCA
CAATTTCGGG CACAGCTTTG TAAGCTCAGT GAACTATGTT GACCATGATT
TCGTTGCCGA CGAAATTGGG AAGATTCACA AGAAAGGCTA TGATATTGAA
ATAAACTGGC TTACAAGGGA GTTCAAGCCC GCTCAGCTTG AGTCAGAGAG
AATAAAAAAA TCAATTGGTT ATTGGGGTGA CAACCTAAAG AAACATTGTG
CATCCCATAG CGTAAATCTG GAAAATCTAT GTTCTTTATC GTTTATCTGG
CCGACAGGTC AAAGTAAATA CATGCATGCC ATTGACGACA AAGGCACAGA
ACACAAAATT TACATCAATG AAGCGCAGTG ATACGCATAT AACAATTGGT
TCAAACCGTT CGCTGCGCTC ACTGGGACGG GCTAAAGCCC GCCCCTTAAC
CAAACGTTAT GCGAGCCACC ATGAATAGCG AAGAATTATA CAAAAAGGCT
ATGGAGTTAG AGTCCAAATG CGAGCATAAA GCGGCAATTT CAACTTACAA
AGAAATTGTT AAGAAATCTA ACGATCCTCG ACACTTCATC GCATTCGGAG
TTTGCCTCCA AAAATGTGGT CACTGGAAGC AATCCATCGA GGTATTAGAA
TCAGGAATTG CACTGAAGCC TCACTATTGC GAGGGTGATG CTCGTCTATT
TTTAGCAAAA GCACTTTTTA AATCAGGCAA AAAAGGCCTT GCGATAAAGC
AATGGCAACA TGTATCAAAA ATGCAACCTG AGTACCCAAG TTATGAGTCT
GTGCAAAATG AAGCCAAGAA AATGCTTGCA CAAAACGCAT AACAATTGGC
TCAAGCCGCT CGCTCCGCTC ACTCGGACGT CCGTAAGCTA CGCTTCCGGC
CGCCCCTTAG CCAAACGTTA GGGGCCAAGA TGGATCTTCG CCAGACAAAG
CCAATACTAG TTACAGTCTT AGCCACTGCC TTGGTGCCAT TGGTTTTTGG
CTGGTATGCG TATTGGGAAA ATCCTCAAGG CATACTTTTG TACACTCCGG
TGGCCGGCCA TCCCCATCCT CAGGGCTCTC CAGCATTTCC TATTGGAGTA
ATGGTTGGGC TGGCCGCTTC ATTTCTGCTC TCTTTGCTTT TTGTAGGCCT
AGGGGGAATC GCTGCATACA TAGCAAGTTC AGTGAGCTCA AAGGCTAGGG
CTAAGCTGTT TTGCAAAATC GCAGTCACAT CCCTGGCTAC TTCAACTATA
GGAGCTGCAG TCTATGCAAT GCTCCCCTAA CAAATGGTTC AAAGCCGTTC
GCTTCGCTCA CTCGGGACCG GCTAAAGCCG GCCCCTTAAC CAAACGTTAG
GCAGCACATA TGACTCGTTC GTGCCTATAC ATGTTTATCG CCTCAGCCTT
GATAGCGTGC GGCGATCCAC CTCTATTGGT TACGCCACTG CCAAATGGCT
ACAATTTCCA TTCCAACGGC GGGGAGTTTG GCTACATCAA GAATCCAGAT
GGATTAAGGC TCGCCGAGTA CTTTGGTATT CGTAATGATG GTCGCGAAAC
CTGGTGCACT GACTTTTCAT GGGAAAGCGA TATCGTCATT TGTAAGCTTA
TTGAATATAG CCAGCATGGA TTTGACGCAT CGCATACAGA GTTTTCTGTA
CTTGACACAA AAACTAGCGA GGTTAGGGTA TTTCCCGATC AAGCGTCTGC
TCAAAATTTC TGGGCCGCAC GCTTTAATTC AGGACTACCT CAGCTTCACC
GGCACTACCC TTCAACCTCA GAGAAGTAAT ATTTTGTGTG TCAGTGCAGC
CTAACAATGC GCTCAACTGC CGCTCACTTC GTTCGCTGGA CAGTCAAAAG
CTGCGCTTTT GCCTGCCCGT TAGCTTAATC GTTAGAGGCT TATTTAGCTC
ATGCGCATAG ACATAGACTT TTCAATATTC ACGCTCGCAC CGTCGACCGA
AGGCGTAATA TCAGGAAAAA TCGAGGTCAG TGAACTACCT AGAACTGGCG

AGATAATTTC ATTCTCCTTT GCGCCAAACA AGTCTAAATT CCCGGCAGAG
CCAAGATTCA ACCCGTTGCT TAAAGTTGAG AGAGTGATTC ATAGCGTAAA
TGGTCAGAGT CCAGCTCTTC AGTTAGAGAA TCTGATGCTA CCAAACAGAG
AAAGTGTCGC TGAAGTCACT GCTTTCCTAG AGCAAGGCTT TGGCCTATTT
TTCAGCCCAA CCGGTGAGTA ATCCTCTAAC AATGCGCTCA AATCGCTCAC
TACGTTCGCT GGGACCGGCT AAAGCCGGCC CCTTAGCTTA ATCGTTAGAG
GTCAGCACAT GGCAGTGCAG CAACTCGGGC CAACCACAGT ATCCGTAACC
GAATTTGCAT GGGACGGAAG CGATCTTGGA AATACTGAGG CCAATGAATT
CTGGTCACAG CTCTCTGCTC AGCTTCAAAA AATAGCTATC TCTGAGTTTT
TAGCTGGCAA TCGCCCCAGC AGCATTCTTC GCAACGACCC ACGAAACATT
ATTGTTCTCT CATTTTCGGC GCCGCCAAAG TTCATTAAAA TCAACCACTG
GCTCTCTGCG TGTACACACA GAATTTCAAC ACGGAAATTA CTGCTACGAC
GGAAACGGCC TGTACTTACG AAAATTTAGA GTCTGGCGAC TTTCTTGCAT
TCGACACAGC GGCGTTGGTG CATGCCCTCT AACAATTGGT TCAAGTCGTT
CGCTTCGCTC ACTGCGGGAC CGGCTAAAGC CGGCCCCTTA ACCAAACGTT
AGGGCACCGC GCATGAGAAA TGAAGACGGA ACCTTTTGCA AAGACTGCCA
CCATCAACTT GATGAAACAC TAGCATCTAG CGCAAATTAC TCATGCCCCA
ACTGCGGCTC CACAAAAAAA TACATGAACA TGTCCATCAC TGATGGAATT
GGCCTATACG ACTCTTTGGG TGCCCAAGCT AAAGATCCAA GTTACCCGGC
AAAAGAAAAA TCAGATGGGA AACATTTGTT GGCTGGGAAC GCAGTCATAA
ACTGCAAAAA ATGGTTTACA AGACAAGAAC TATCGATCGA ACCAATGACG
CATACCAAGA AATAGTAGTC GACCTTAAAA CAGGGGGAAT AATTCATCAC
TGTGAAGAGC CACTTTCAGA GCAYTTKGGC CATGGCACCG CAAAACCAAA
GCCCTAACAA ATGGTTCAAG TCACTCGCTT CGCTCGTTCG GGACCGGCTA
AAGCCGGCCC CTTAACCAAA CGTTAGAGGT TACCTGTGAC AGATTCGCGC
CCGTTACTGA TCCCTGCCTC GCAATATGAT ACGAGCGTTC TTCTCGCCGA
ATGGCAATGG CTCACCCCCA AAACGGATAC GCCACTTTTT ATTTCCATAT
TCGGAGACTG GGTATTTGGC AACCCCAATG GAAGTTTGTG GGTTCTTTCA
CTCCTAAAAG GCACTTACGA GCAAGTAGCC GCAAACTCTA ACGAGTACAA
CACCCTCAAC AAATCGGCGG AGTGGATTGA TCAAACATTC ATCGCCAGTT
GGCAGTCTAT TGCCGCAGGC CATGGGTTAA TCCCAGAACC AAACCAATGC
CTCGGCTGGA AGGTTCACCC ATTATTAGGT GGAAGTTTTG AGCCAGCCAA
TCTCCAACTC TTCAACATGT CGGTGTATCA ATCGCTTATG GGTCAACTTC
ATCGACAGCT TAGCCAAAAA CAAACCCCGG CAAGTAAAAA ACCATGGTTC
CAGTTCTGGT AACCTCTAAC AACTGGTTCA AGTCGTTCGC TTCGCTCACT
GCGGGACCGG CTAAAGCCGG CCCCTTAACC AAACGTTAGG CGCAAGGGCA
ATATTGGTTA TTCAGCACCG AGCCAGGGAA CACAATCACC GCATCAGCGC
AGTGTTCCTG AATCGAATGG TCGCCTGACA GTAGAGGCCG TTATTTGTGG
CCAGCAAAGG AGTTGCTTTC AAAGAATGTA CACGTCACAA ATAACTTCCG
GGGCCAAAAC CGACACGCCG TGCGCACCGT CGGTCAAGCG CAGCGCTGGC
CTCACTTGCA GCGTACGGCT GGGCAATCTA ACAATTGGCT CAAGTCGTTC
GCTTCGCTCA CTCGGGACGT CCAATAGCTG CGCTATTGGC CGCCCCTTAG
CCAAACGTTA GGCCAACATA CTCAACGCAT GAAAACAAAA TATCACATAA
ATATAATTAT ATTTCTCGAA ATCATAATTC CTTTAGCACC AATAATTTGG
GCAATTTTCA CTCAGTCAAG CCCCGGCTTT GGCCCAACCC TTATATCAAT
GCTCATCCTG CACATCGTCG GACGAATAAT TAGCCGAAGC ATCCCTGCCA
GCTGTGACTC ATGTGCTGAA AAAATAAAAC CCAAAGGAAC CTCCGCAATC
TACTACAACT GTCAAAAGTG TGGATTTAAA TACTCAAAAA CACTTAACAG
CAGCAAAAAC TTCCATAACC ACTAACCAGA AAATCACTAA GGCGCCATCA
TGTTATAAGC GCCGTAAGCA CTAAAGACTT GTACAAGCCT AACAACTGGT
TCAAGTCGTT CGCTTCGCTC ACTGCGGGAC CGGCTAAAGC CGGCCCCTTA
ACCAAACGTT AGGGCACTCA ATGCATCGCT TCCTAGCCAC ATGCCTACTA
GCTACATCTA TTAAGGCATA CGCAGAACCT GAAAATAATA TCGACTGCAG
CAACGCATTC TCAACGCCGG ACATTGAACA TTGCGCATCA ATCTCTCTTG
AGAAAACAGA GAAAGAGCTA AATTTAGCAT ATCAAAAATT AGTCAAAGAC
CTTTCTCAGC CAAACAATGA ATACGAAAAT TTCACCGAGT ACAGGAAAAA

ACTTTTAACG GCTCAAAGAG CATGGATCGC GTTCAGGGAA GCAAACTGTG
CCACTCAGTA CGAAATGCAC AGATCTGGCA CTATTCGCAA CAGCATCTAT
CTAGCCTGCA AAGAAAAGCG TGCCAAGCAG CGAATAAAGG GAGCTTCAAA
ATTATGCTCC GTACTAGCCC TAACAAATGG TTCAAGTCGT TCCGCTTCGC
TCACTGCGGG ACCGGCTAAT GCCGGCCCCT TAACCAAACG TTAGGCCGAC
AATCGCAATT CCTAGGACTG CACGTGAACT GGATCCGCAA AATGTTTCGG
CGCACAGCAC TAGCGCCGCC CCAACATCGC GAGGACGAAG CTGTCAGTAC
AAGCCAAGAA GGAACGCCTC CCTTTCGTCA TTTGACAGTT GAGAATTCAT
GGGGAAGTTG AGGGCGGAGC TATTCCTTCA GTCACCACCC GAGAACATCC
TCAGAAGATC TGTTTCTCGT TTGGCGTGCC TAAGTTCGGA TGGTCAACGT
TCGAGATCCA TTTCGTCGGA AATGGCCACT TCATCTGCGG CATCTCTGAC
ACTCCAAATG ACTTCTACGG TGACTTGGCT ATCGCCCTGG CTGAGCAGAA
AAGTTCTTTT TCGGTAGCGG CGCACCTTGA GCCTGAGACC TTTGCCTTCT
ACATCGTTGA TTCGACAATG TACTTGTGCA AGTTCGATGA ATTCGACGAT
TATGAGTCCG CCGCCGAAAG CCACGAACAG TTGGTCTCCC ACAGCTTTAT
GTCCATTGAA GTATCTAGGG AGTACTTTCA GAAGTCTCTC AGGACCTTGG
CCGTCCAATG GCCGGATACG CCTTCAAGAG ACTGGGCGCA CCCATTTCCA
CGTGCGCAGA TTGAAGGCTG ACTGCCTAAC TATTCGCTCA AGCGGGCAGC
GTTAGGCGCC CTCATTCGGA GTCACGCTAT GGCAACCCGA GAAGAAACAG
AAGTAGCCAT TGCTGCTCTT CGCAGCGAAC TCAATGGCAA CGAATCGGAA
TACAGCTTTC ACATTCCCGG TTGGGCGCCA GAAACATCAG TCATGGGATT
TCGCTGGATG CAAAGCCAAC TGTGGGAAGG CTTCTACGTA AGCTATCGCG
TAGAGCACTC GGCCAAGCGC GTCGAATTCA AGTGCTGGGA GTACGGCGAG
CCCGAGCCGT CTTGGCTGCA AGTTGGCTAG GGGGCCGGCA AGATGCAATC
GCGGCGAGCG CCTAACACTG CAGTCAACCG GACACCAAAC TGTACGCAGT
TTGGTTCCCT CCGCTGCGCT CCGGTGCCGG TTACTTTCAA CGTTAGGCAA
CTCAGATGAG TGCTCCAGAC GCAGAACTTC TCGCATTGTT AGCCTACCGA
ATGGAAGCTA TTTCCATTGG GCATTTGGCA TTACGCCATC ACATGACGTG
GGACGAAACA CCTTCAATGG AGGTGTACTT CAATGGCATA CAAGTACTCG
AGGGAAAGGC CACGGGTTTC ACTAATGCAG CCATTGAGTC CGCAATTATT
CATTGCAGGG CAATCCTTGG AGTTTGTTGG GCTGCAGTCC TCCAGACACT
CTTCCACAGA AATTGCAGAG CGCACTCGAC GCAACAATCC CGATGACTAT
GGCATTGAAA GCTTCAATGG CTTATCAATG CTAACCAAGG AAAAAGCACT
AGCCTACTAC TCTGGCGAGC TGCCAGAAGC GGAAGTTGCT CTAGCGCTCA
TATTCCACTC AGCGAACAAA GGGCTTGCAC ACACTACAGT GTCCTTTACG
CGTGACAGTG GCGACGCCCA CCTGATGGAA ATTGCATTTC GCATCGTACC
AATCCTGCTT GTAAATGGCT TCTACGCTCC ACTGGAAATC ACGCCACCAA
AATATGAACT GATTTCACGC CCAAGAGTCG CCATAACAAA TGGTTCAAGT

Figure 3D-1
SEQ ID NO:4

GCCGGCCTGC AACAGAGCTT CAAGGCCGCC GGTGTCGGCA TTCGGCGACC
GCGTGCCGCC TTTATGCCAA CACCTGGCAG ACGGTGGTCG GCATGTTCGC
CACCGCCAGC TTGGCGCCAA CTGGTCGTCC TGCTCGCCGG ACTTCGGCAC
CCAGGGCGTG ATCGACCGTT TCGGCCAGAT CGAACCCAAG GTGCTGATCG
CCGCCGCCGG CTACCGCTAC GCCGGCAAGA ACCTCGATCT GACCGCCAAG
CTCAACGAAA TCCTCGAACG CCTGCCCTCG CTGCAGCAAC TGGTGGTGGT
GCCCTACTCC AACCCGACAG CCGGGGCGGG CGACTTCCGC AGCGCCGCCC
GTGTCAGCCT GTGGCAGGAC TTCTACCAGG CCGGCGGTGA ACCGAAGTTC
ACCCCGGTGT CCTTCGAGCA GCCGCTGTAC ATCCTCTATT CCAGCGGCAC
CACGGGCGTG CCCAAGTGCA TCGTCCACGG TGTCGGTGGC ACCCTGCTGC
AACACGTCAA GGAACTGGGC CTGCATACGG ACCTGACGGC CGACGACACG
CTGTTCTACT ACACCACCTG CGGCTGGATG ATGTGGAACT GGCTGGTCTC
AGGGTTAGCC TTGGGCGCCA GCCTGGTGCT GTTCGACGGC TCGCCGTTCC
ACCCAGGTGC CGAGCGCCTG ATCGACCTGA TCGACGCCGA GAACATCAGC
CTCTTCGGTA CCAGCGCCAA GTTCATCGCC GCCCTGGAAA AGGCCGGCGC
CAAGCCGCGC GAGACGCACA GGCTGCGCCG CCTGAAGGCC ATCCTCTCCA
CCGGCTCGCC GCTGGCCCAC GAGAGCTTCG AGTACGTCTA CCGCGATATC
AAAAGCGACG TCTGCCTGTC CTCCATCTCC GGCGGCACCG ACATCGTCTC
CTGCTTCGCC CTCGGCAACC CGACCCTGCC CGTGTGGCGC GGCGAGCTGC
AGTGCAAGGG CCTGGGCATG GATGTGCAGG TGTGGAACGA GGCCGGCCAG
CCAGTCATCG CTGAGAAAGG CGAGCTGGTC TGCGCCCGCC ACTTCCCGTC
GATGCCGGTC GGCTTCTGGA AGGACGCCGA TGGCGAGAAA TTCCGTAGCG
CCTACTTCGA CACCTTCCCC GGCGTCTGGG CCCACGGCGA CTATGCCGAG
ATCACCGAAC ACGATGGCCT GGTGATCCAC GGCCGCTCCG ACGCCGTGCT
CAACCCCGGC GGCGTGCGCA TCGGCACTGC CGAGATCTAC CGCCAGGTGG
AGAAGGTCGA GCAGGTGCTG GAGTCCATCG CCATCGGCCA GGACTGGGAA
GGCGACGTGC GCGTGGTGCT GTTCGTGCGC CTGCGTGACG GCGTGGCGCT
GAGCGACGAA CTGCAGGCAC AGATCCGCCA GGTGATCCGC GCCAACACCA
CGCCGCGCCA TGTGCCGGCC AAAATCATCG CCGTCGCCGA CATCCCGCGC
ACCATCAGCG GCAAGATCGT CGAGCTTGCC GTGCGCAACG TGGTGCACGG
CAAGCCAGTG AAGAACACCG ATGCCCTGGC CAACCCGCAA GCACTTGAGC
TGTATCGCGA TCTGCCGCAA CTGCAGTCAT GAGCCGGTAA GCGACACCGT
AAGGGCAATG GACTGCCACT CCAGCATCTA TAGTGGATGC GCATAACCCG
GACAGGATGT TGCTGATGCA GGTGGTTTTC TGCCTGCTGG TTGCCTTGCT
GTATGTCGGC GGCGTGGCCG CTGACGAACC ACTGGCCTTG CATATGCCGG
ACGCCCCGCC GCTGACCCTG TACCACGACG AGCGCGGCCA CGGCATGGTC
GGCGACATCA CGCTCGCGGC CATCACTCTC AGCGGCCGAA CGGCACGCAT
CGTCGACGAG CCCTGGGCCA GAGCCCAGGT GAACGTCGCC AGCGGCCAGA
ATCAACTGAT CATCCCGCTG TCGCGTACCC CGGAGCGTGA GCAACGCTAC
ACCTGGATCG TCCCGATCAT GCCGCTGGAG CGCGCCTTCT TCAGCCTCGA
CAAACCTGTC AGCAGCTTCG CGCAGGCACG CCAGCGCTAC CGGCGTATCT
GCGTCGGGCT CGGCACCGCT CAAGTGGAAA TCCTGCGGCG CGAGGGTTTC
GCCGACGAGC AGATCATCCA GCTCAAACTG GGCGAAAACC CGGCCATCCT
GCTCGAACGC GGGCGTCTCG ATGCCTGGTT CACCGGGATT CCGGAGGCGC
TGTACATTTG GCACAAATCT GCGGAACAGC GCCGCAAGCT TTATCAGAGC
CCGGTCCTGG CCAGCACCGA CCTGTACCTG GCCTGCTCCA GGATCTGCGC
CCCGCAGATC GTCGAGCAAC TGCGGGCCGC CGTGCTGCAA CTGGAGGCCA
GCGGCGTCAG CCCGCGCCTG CGCCAGGCCT ATCTACCCGA GCTCGATCGA
CGGTGAGCAC CCAGCCCACG CTGCCAGGCC TATAGAACAG ATCATCCACG
CCGCAGGCCT GCGCCGGCTT GAGCATAACG CCAGCACTGG GGCAGACTGC
GACTACCCCG CACAACCCCG GTGACAATAT GGACGTTGTC AAAACCCTCA
AGCCCGGCAA ACCCGGCACC AAGCGCTTGC AAGAACGCTA CGGCGAGCAA

CTCGTCGCCG TCCGCTACCG CCTCGACCGC AAAACCAACA CCCACTACAC
CACGGTCGAA CTCATCGTCG AACAAAAGTA CGCCCTGTAC AAAACCCCGC
CACCCGCTCC CACACCTCCG GTAGCCCTGC GCATCTTCCG CCACGAAAAC
GACCTCCAGC GACTGATCAG AAGCGCCGGC GGCAAGTGGG ACCGTGAGAA
TCAGGTGTGG CTGATCGAGC GAAGCGAGGC CGAGAGGCTG GGGCTGGCGG
AACGGATCAT CTGGACATAA TGGCTATATG TGGACATCAA GATGCCTAGT
AATAGCCACA AACACCCAGC ATCGGACACT ATGCCTACCC CTAGGCATGT
CGTATAAACA CTAGTTATAC AAATATCATA TGAACGACGC GACCCTAAAG
CTAGTTAATC AAAGACAGCT CGTATCGGTA ATGAATAAAA CGAAGTGGAC
TGAGCTGTGC AATTCATTTG ACTGCGAGAA TAAAGCATCT CCGAATGTTC
GCTATAAATT AATTTACAGT GAACAAGAAT TCGGTTTTTC AAAAATATGG
TGGAATCAGC TTTTGCATGA GTGCGAAGCA ATCGAATGGA TTGATTTCAA
ACTAGTATTG CGAGAACACC GTGGCAATCT ATTGCCAGAC AAAGAAATTG
ATATAAGCAA ACAAATTAAG GAAGCACTAC AGGCGCATGA CATCCCTTAC
TCTGTTGAAG GAGAAAATCT TAGGGTTTGG GGCTATATTA GCGCAGAAAA
GAGTCCAGTA TTCGTATAAC AATTGGTTCA AGTCACTCGC TTCGCTCGCT
CGGGACCGGC TAAAGCCGGC CCCTTAACCA AACGTTAGAT GCTTATGAAA
AAGACAGTTC TCATACTCGT CCCAGCATTA CTACTCTCAG GATGTGGCGA
CCCTGAATTT CACTACCAAA ATGGTGACGA ATCAAAAAAT ATAACGCTAC
GCATCCCTAA GAATTACATA AATTATTTCC CTGGCGTGAA GTACGAAAAA
GACGGACCTG TCGTCATCAG ATTTTCATAT CCACAATTGG AGCCACTGAC
AAAAGCCCTA CCAGAAGAGC AAAAAGTAAC TGTCAGCATT AGTCATTTAT
CCAGCCTGGA ACTCACCACC CAAGAAACCA GAAACCCCTA CTGCGAAACA
GATAAAAAGT GGAAACTCCT ACAGGCGGCG GGCATTCACG GAGAGTTCTA
TAAATTCATC GGAAAATCTC CCGGCAGCGC CAGTGCAGAT ATAACTTATA
AGCCCATCAA AAAAACACTT GGCCTTTACT GCATTACATG CGTGGAAAAT
GCAAATTGTG AAATTCACGC AGTATCTAGC CAAGGAATAA GCTATTCCGC
ATTTTATACA GAAGACTTAA TGCCAGATAA GTGGCACTCT ATTTACATGG
CAGTCGACAA AATCCTTAGC AAATTTACAG CATCGTCGAA AGGCATCTAA
CAATTGGTTC AAGTCGCTCG CTTCGCTCAC TCGGGACCGG CTAAAGCCGG
CCCCTTAACC AAGCGTTATG CAAGCAGTCA CCCATGAGGA AAGCACCCAT
ATGGAGCCAG TATGAAATTG AGCGACATAA GAGCTCTAAT CATTGAGTCG
CCAGGATGGC GAACAGTATT TGCATTTATT GTCCCACTAA TCGCAGGGAT
TCTGTCGGGA ATATTCGTAT CAGAAATAAC GCATAGCTCC GAAATTGTTT
GGAAGGAATT TTATAAAGCA AAAAGCTTCT ACGGGCTATT GGCTTTGAGC
TTGTGCATGT ATTTTTACAA TAAAGCCATT TATCTACATG AAAGAGAAAT
TTCTCGCTTC CTAGACGCAG ATTACTGCAC CGCTTACATG AGAAGCAAAT
GCCTGCCAGA GGCTGCAGAG CGATACAAAA AGCTTATACG CTCTGGCGAC
GGCGGCGAAT TGAAGCAAGC AATGGATGAA CTGAAGAAGG TGCTCAAATG
AAAGTACTGG CCAGCCCAGA TTTTAATGCA AAAGTGCCGG CACTAAACAC
AGAAACCATT AGTAGCCTTT CTGCATTCAT ATCAAGCGCA GAGCAATATG
AAAAAAATGA CTTCATATTG AAGAATGTAA ACTCAATGTC TCTTCTTGAT
GGCGATATAT ATAGCGCAAA AATCAACTCA AGCAGACTAT ACTTCACCAT
CGGAGCTGAT GAGCAAGGCG ACTACTTGCT GCTATTAGAT ATAGCTGCCT
TGCAAACCGC ACCATCTGTC AAAAGTAGTG CTTTCTTCAC AACAAACAAC
CCAAAAACCA ACAGCTCACT CAATCCGAAG CTCAACTCTG CAATCAACCC
AAAGCTAAAC TCAGCAATAA ACCCAAAATT AAATTCGGCC ATAAATCCGA
AGCTGAACTC AGCAATAAAC CCAAAGCTAA ACTCGGCCAT CAACCCGAAG
CTGAATTCAG CAATAAATCC GAAACTAAAC TCAGCAATAA ATCCAAAGCT
AAACTCAGCA ATAAACCCAA AACTAAACTC CTCACTAAAC CCAAGGCTCA
ATCGAAGCTA TGGCGGCCCG TATTTGTACG ATGCGAACCT TAATCAAGAA
GCGTACTCAG TTAGAGCCAA TAACAAAATC GAAATCCTGT TCAATTCGGG
CGGAGATTTT TATGGCTTTC TTGTAAGCGC TAACGACCGA GTGAAGATTG
AGTTCGATAC AGGAAATACC TGGACAGGTT ATTACGTTAA AGCCAATGAA
AAAGTTTGGC TTAGATATTC GCTTAACAAC GAATGGTTAG GGCTACTTGT
CTAGCCCGCA TAACAAGTCG CTCAAATCGC TCACTTCGTT CGCTGGGACG

GGCTAAAGCC CGCCCCTTAG CTTATCGTTA GGCAAAAAAA TAGCAGGCAG
GCTCAGTAAT ATGAAGTTCG ATAGAATAGC TCGTGAAGCG TTTGGCTCAG
TGCTTGGTCC ACTGGGGTTC AGCTGTAGTG AGTCGAAGGC ATGCACCTTC
TATAAAAAAG TCGGCACTGA GCTCTATCAT TTTGTCATGC CAGATCAATT
AAGCGGCCAG GAAAAGTATG ATATTAAAGT TTTTTTCCAC TCGCCGCTCT
TAGAGCCAAC CGCATGGAAT GACAAGTTTC CGGACACCTT GGGGATTCCC
ACAGATAGCT GGAGTTATCT TTCTAGCCGT ACTGGCGTTG GTCCACGACA
AGAGCTGTTT TGGTGTCGAA CAGAAGAAGG ATTTATGCGT AACTTTGAAT
CAAAGGTAAA GCCCGCACTA CTTCAATTTG TAGCCCCATA TTTTGATTCT
ATACAGACAT TGGAAGAGGC TATTCCACTA ATCAAGAGCA GGCACTATGT
GGCAGTGGCG TCTACGCTAA ATGCTAACTA AGCAATGCCA AGGTCTTCCA
CCGGCACCTC CGTATCGGCC TTGACAGATA GCAGCAATGA GTTTCCAGCA
AAAACCAATG CGCCGCTTGC AAGGCTGTTT CGGGTTAGCC ACAGTGCGGT
ATTCATTACC TGCGTCCGAC TCGATACCAA TTGCCTAACA ACTGGTTCAA
ATCGCTCGCT CCGCTCGCTG GGACCGGCGA AGCCGGCCCC TTAACCAAAC
GTTAGGCTAC ATATGAGAAT CAGCGCAGAC CAGCTTGCTC AAGAATCACT
GACTGAGTTC GGCGTGCTGG CGGCTAAGCT TCTGGCAACG CGAGAGCTTA
GCCAGTTGTC CGAGAAGTTT GGGTATGCAC TGGCCTTCGG AAGGGAACCG
GCGGCTGCCA TAGCTGAGGA CCTTGCTAGG TGCTTGTGCG GACAAAATGC
TTCGCCGGCA TCTGAATACC CCAAAATCAC CGTTAAGTAT TTCAAGGAAA
ACGAAAGTAG TCTGTTGGCA CTCGTAGAGT GTTATGTACA AATGACCGCA
AGCGCAAACA TTCTTTTAGA GCTGGTTGCC GCACGAAATG GAGAGGCAAT
AAATCTGTAT CTAGAAGGCT TGAGTGTTGT AGCCTAACAA TGCGCTCAAA
GCGCTCACTT CGTTCGCTGG GACCGGCGAA GCCGGCCCCT TAGCTTAATC
GTTAGGTGCC TCAGGAGGGA TCATGTCTTC CACAGAAAAC AATAGTGATG
ACTGGCGAGA AATTCGAGCA AGAGCGGACT CTATCGCTAA TGCCATTTTC
CTCATTTCTG GCGGGGCACT TTCACTTTCA ATCTCAGTCA TCCTCAGCAA
CAAAAGTGCC GGGTACATCA CTGCACAAGT GGCATGTATT GCGTCCCTCG
CTTGGTACTG CTTGCTGGCG TCACTGATTC TCTTTCTTGC TCTTAAGGGG
CATATGATTC TTCAGGCATA CCTCCTACAA TTTCGCCCAA ATTACGTCAA
TAAACATCTT AGATTTCTTA ATGGTATAAG CTGGGCCATT GGATTAACCG
GGTTTATTTC CTTCATTGCA GGCATGTTTC TTATGGTTCG TACCGCAATA
CTTGCCGTCG GCACCTAACA ATGCGCTCAA CTGTCGCTCA CTTCGTTCGC
TGGACAGTCA AAAGCTGCGC TTTTGCCTGC CCGTTAGCTT AATCGTTAGC
GGTCATAAGT ATGCAGATTA ATTTCTATAT GGCAGATGAA GATCGAAGAG
CGTTCCACGA ATACCTATAT TCTCGTGGCG CATACCTCGT TCCGGAGCGT
TGGCCAACCA GAGATATTCC CATAGTCCAG GCGGCCTCCG AAGAGGCAAG
TGAGTGCAAA GACTTCAAGA TTTTCAAGTC TGACCTCTTC CCTCAGTCCG
AATTTCAGAA CAGGGCTTGG ATAACGTGGC ATGAGCCAAC GAAAAGGTTC
TACGTTCATG GGCCTGGAAT TCAGTATCTT GTATCGTTCA CTGATGCAAA
TGGAATCCAT CGTGGCCGCC TCTATATGGG CCTTGTTTCT CCGCGTAGCT
TTGTTGAGCC CCACGGGCAA TCAGTTGATT GCTACGCTGA AAACGAGAAA
AAGTACAAGG CGTTAGAGAA TTTCCATAAG AGTTGCGCGC GCTATATACG
CAATCACTAC CGCAAAGATG AGGGTGGTTT CTACCATGGC AAAGCAAGCG
ATATGGCTGT TCAAAACTAC GGCGTCTCAA AGACGCAGCT GTGACCGCTA
ACAATTCGCT GCAGGCGCGA CGGCCCTGAC GGGCCGCGGC CTGAGCTCAA
ACGTTATAAC CTACAAGGAA GACCAAAGTA TGCGCCACCT AGCAATAGCC
CTCTTAATAA TGTTCTCTAC TCAAGTTCTC GCCGACGGCA AGAGCGAAAA
GATAGAGAGC CTAATGAAGG CACTTGGACT AGTAGACACA TGGACACAAC
AAATTGAACA AGGAAAAATT TACAACAGAA AGATCAGCTC TCAAATGCTG
GATCAAATTT TATCCCAGCT GAATCCAAAT GAAGAGTTTC AGCAAAAATT
CAAAAAGGCT TCAGATAATT TCATAACAAA AACAGAATCT CCATGGTCTC
CAGAAAAAAT TGTAGAGGTT TGGGCTAGTT ACTATGGCCC AGAATTCACA
GAGGACGAGC TTGACCAATT AATTGCATTT TATACTTCCC CTCTTGGCCA
AAAAGACATC CGTGTTACTC GCAGTTCAAT GGAAAAATTC TCGAAATACT
TCCAAGAGGC CGGGCAACCA ATACTAGAAA AGGCCACCGC AGAGTACATT

CAGGAAATGA AGCTCATCGC CAAGGAATGT AACTGTACCA AGTAGCTTAT
AACAATTGGT TCAAGTCGTT CGCTTCGCTC ACTGCGGGAC CGGCTAAAGC
CGGCCCCTTA ACCAAACGTT AGGCACTGCT ATGGCCTTGG TCGAGTACGA
ACTGATCATC AATGCGCCCC AGACGGCTGT CTATGCCGCA TCTCAGGACT
ATTCAGTTAG GTACCAGTGG GACCCCTTCC CTGAAAAAAT TGAACTCCTA
GGTGGTGCAA CCGAGGTAGG AATTGGGGTT AAGACACTTG TAGTCGCCAA
GTCTGGCTTA ACAATGGAAG TCGAGTTTGT TCAGGTTGCT CCTCCTACAA
CGGCAGCCAT AGTCATGACC AAAGGCCCAG CATTCATCAA GAGCTTTGGT
GGTAGCTGGG TTTTCAAGCC CATCACCGCA AACTCTACAA AGGCAAAATT
TCGCTACTCC ATAAAAACCA AGAAATGGGC AATACCCATA ATCTCAGAAT
ACGTAGCAAG TCTTTATTTC AGAAGAGCAG TTAAGGCCAG GCTTGCCGGT
CTTAAAAAAT ACTGCGAGCA AGGCGCCTAA CAAATGGTTC AAGTCGCTCG
CTTCGCTCAT TCGGGACCGG CTAACGCCGG CCCCTTAGCT TAATCGTTAG
GCTGGCCGAA GATATGAGTT ACAAGAGATG GATTTGTGTC CACTGCGATA
CAGCCAACAC CACAGCAACA GATATTTGTT CAAAATGTCA CAGATCCAGC
TATGAAGAGC CGGCAATAGC TGAAACTCCA ATAGCTAATT CTTACCAAGG
CATACAGCTG TTAGGCTCTT GGCTTTTTAT CCCACTAACC CCATCCATTA
TGGTAATTGC AATAAGGGAT GAAGTCTGGT GGTTCGTCCC ATTTGGGATC
GCAGTTATTG CGCTCACAAT ACTAAGTGAA AAATCTAAAT TCTTAATTTC
CAATACTACT TGGTTCAAAA ATATAGCTTT ATTTTATACC CCAGCAGCGG
GTGTGCTTTT CCCTCTTAGC GTTTTTCTCG GAAAGAATTG GGCGGCCGCA
TTTATGGCAA TGCACGTGGT TGTTCACCTA CATGCTGCAT TTAACATGCA
CGCACACTGC CAAAACCACA AGCCAAGAAA TGAAAATTAA CGATAAGTAC
TCTAGGCCTA ACAATAGGTT CAAGTCGCTC GCTTCGCTCA CTTGGGACCG
GCTAAAGCCG GCCCCTTAAC CAAACGTTAG GTTCCATCAT GTGCTACATG
GCCATAGTAA GTACAACATC CGAATCTGAC CTAACGGCAC TAAATACGCC
CCTAGCGCAG TTCTCAAGGA ACGTAAACGC CATACCAGAA GCGGCCCTCC
TGCGTTACCC AAATAAATGG TTCCTTGGCT CAAAAGACGG TTGCAGTTGT
GCATTCAGGC ATCTTGATCA AAATGCTACA GATCTTGGTT TCTCAGAGCC
CGTGGACTGG TGGGAAGAAG ACCAAGAAGA TATAGATGCT ACGCTTCAAG
TAGTAGAAGC ATTCCATACG ATATTGCGCG ACGGCCATAA ACTTGACTGC
ATAGATGCTT GGGCCAACGA TAGCAAGGAA CCTAAAAATC TTGCAGGTGA
TCTTGTGGTT GACCTAAATA AAGTTGGCGC CAAGAGTTTC CGTTTTTTTG
AAGGCTATCG CTTTGAGCTC GAAGCCAGAA CCTAACTAGT GGTTCAAGCC
GCTCGCTTCG CTCACTCGGG ACCGGCTAAA GCCGGCCCCT TAACCAAACG
TTAGGCCCTA TATGCAGTAC TCAATTGCTG ATACTGAGAG TTTTCATCCC
GTCATGGATG CTGAGATCAA GGGGCACTGC GAATTCCCCG TTGACCTAAT
TTTTGTCCCG GACATTGCAG AGTGGGCTGC TTCCAGATGC GGAAATCTCA
TCGGCAATCC TGTCGCAATG GCAGTTAGAG ACGGAGCCAC TAAGGGGGCA
GGCATTCTCA TTAGGCAATC AATAGACGAA TCGCAAGTCG ATAGCATCCT
ATCTCGAATG GAGTTCGGCG GCTTTGATCG TGCGCGGTCC ATACTGTCCT
CACCCGAGAA ATTTATGCGG CATCTAGTCC TCCATGAGCT CGCGCACTTG
ATTAACAATT GGGGGCAAGA CCGAGAAGAC GACTGTGATG AATGGGCTTT
TAAGCGTCTT GGTGCTAGGG CCTAACAATG CGTATATGGA CTCTCCCCAC
AAGTAGTGGG CAAAGCTTTT GCTCTGCTCC TTTCGTCGGT GCGGTTCCAT
GCGTATATCC GGCCTTTCGC GAGCTCAGTG CTCTGGCCAT TCTGTAGTTC
GCGCAGCGGG GGCCCAGCGC TCAAGCGATC TCCAGGATCA GCATCGTCAC
AGGCCTGCTG TGCAGCCGGC CGAAGGAAGA CCTTCCCTCT GCTCTCCCTC
TCCCCCATCG CTCCTAATGG GCGCGTCGGC AGCCCAGGCT GCCCCAGCCT
ACAAGGAGCA AGAACATGCC AAACCTGACG TTGCATCGCG GTGAGAAGAA
CAACTGGTGG CCAGCACCCG AGGTGCGGCT GATCTGCGGG ATGACCCTGT
ATGGCCCCTG GCAGCCGACG GCGGTCAGCT CGCTCTGGAA CACACTCAAG
AACGAGGTGA AAGGCGCCCT GGGCACCTCG CTGGAGAAGA AGGTCGCCGC
CTACGCCCAA TACCTGCGGG CGACCGGTCG CCCCTTTGCA CTGGCGACGG
CCTGGACCGA AGTCGGCTCC TTCACGTCCG ATTACAACTA CGTGATCCGG
ATTCCCAATG CACATTTCTT CTACTGGGGT GGCACCAAGG ATGCCCCGGA

CCTCGGCGCT GCCGCCGCGT GGACGACGCC CGAGCAGGTG ACAGCCGACT
TCATCGTGCT CAACGCTCCC ACCGTGGCGG CCTCAACCAT CCTGGGCTTC
GGGCACCACA CCGGCACGCG GGAGATCACC TTCTTCCACG ACCTGCCGAT
CGGCCTGATC GAGTCCTGTA ACGGCAGGCC GATCACCGAC TATGCGATCA
AGAGCAAGTC CGACCTGAGC TTCGACGAGA GGATCAAGTA CGCCAAATAC
CTGCGCTAGC CGCGTACCCC GCGTCCGAGA GGCTTAGAAG CTAGGGCGGC
CGGGGTCTTC CGGGGGGGTG TCTTCCTCGA TTTCCTCAAG CTTGAGTTCC
ATCGCCCAGT TGGCCGGTGC CGCCGTGGGC GCGGCAACGG GTGCGGGCGC
CGGGGCGGCA GCCTGCGGGG CGGTGGGGTT GTCCTTGTAC AGCTTGAGCT
TGAGGCGCAC GTTGTTGGCC GAGTCGGCGT TCTTCACTGC CTCCTCCTCG
TCGATGACGC CTTCATGAAC GAGGTCGATC AGCGCCTGGT CGAAGGTCTG
CATGCCGAGG TTCTTCGACT TCTCCATGAT CTCCTTGAGC TCGGAGAACT
CGTTGCGCTT GATCAGGTCG CGTACGGTCG GCGTGCCGAG CATCACCTCT
ACGGCGGCGC GGCGCTTGCC ATCGACGGTC TTGACCAGGC GTTGGGAGAC
GAAGGCGCGC AGGTTGTTGC CGAGGTCGTT GAGCAGCTGC GGGCGGCGCT
CTTCGGGGAA GAAGTTGATG ATGCGATCCA GCGCCTGGTT GGCGTTGTTG
GCATGCAGGG TGGAAATGGC CAGGTGACCG GTGTCGGCGA AGGCCAGGGC
GTGCTCCATG GTTTCGCGGT CGCGGATCTC GCCGATCAGG ATTACATCCG
GCGCCTGGCG CAGAGTGTTC TTCAGCGCGG CGTGGAAGCT GCGGGTGTCC
ACGCCGACTT CGCGCTGGTT GATGATCGAC TTCTTGTGCC GGTGCACGTA
CTCCACCGGG TCCTCGATGG TGATGATGTG GCCGCCGCTG TTGCGGTTGC
GGTAGTCGAT CAGCGCCGCC AGGGAGGTCG ACTTGCCGGA GCCGGTACCG
CCGACGAACA GCACCAGACC GCGCTTCTCC ATCACCGTCT GCAGCAGCAC
CTCGGGCAGC TTGAGGTCCT CGAACTTGGG GATGTCCATC TTGATGTTGC
GCGCGACGAT GGATACCTCG TTGCGCTGCT TGAAGATGTT GATGCGGAAG
CGACCGACAT TGGGCACCGA GATGGCCAGG TTCATCTCCA GCTCCTTCTC
GAACTCGGCG CGCTGCTCGG CGTCCATCAC GCTATTGGCG ATGGCGGCGA
CGTCACCCGG CTTGAGCGGC TCCTGGCTGA GCGGCTTGAG CACGCCATTG
AACTTGGCGC AGGGCGGCGC CCCGGTGGAC AGGTAGAGGT CGGATCCGTC
CTGGCTGGAC AGGATTTTCA GCATCTGGGA AAGGTCCATC GCACGCGCTT
CCATTTGGGT GGAGTTAACA AGGTAGGCCA GCTTTGCCCG GCCGATCAGC
CTGAAAAATG GCGCCATTCT GATGGCGCAA CGAATGCTGG CACAATAGCG
CCATCGCAAA ATGAGGACCC CGTCATGCCC AAAGCCATGG CCCGCCACAT
CCTGGTGAAA ACCGAAGCCG AAGCCGCCGC CCTGAAGAAA CGTATCGCCG
CCGGCGAGGC CTTCGATGTG CTGGCAAAGA AGTACTCCAC CTGCCCCTCC
GGCAAGAAAG GAGGCGACCT GGGCGAGGTG CGCCCGGGGC AGATGGTGCG
CGCCGTGGAC CAGGTGATCT TCAAGAAGCC CTTGCGCGAA GTGCACGGCC
CGGTGAAGAC CCAGTTCGGC TATCACCTGA TCCAGGTGTT CTACCGCGAG
TGATCCAGCG GCTTAGCCGG CCCAGCCGAG GGTAATGGCG GCCAGCACCA
GGTAACGGCC GGTCTTGGCC AGGGTCACCA GCAGCAGGAA GCTCCACCAG
GGCTCGCGCA TCACCCCAGC CATCAGCGTC AGCGGGTCGC CGATCACCGG
CGCCCAGCTC AGCAACAGCG ACCAGCGGCC ATAGCGCCGA TAGGTGTGTT
TGGCCTGCTC CAGGCGTTGC GCGCTCACCG GGAACCAGCG GCGCTCATGA
AAGCGCTCGA TGCCACGGCC CAGCGCCGCA TTTCAACACC GAGCCCAGCA
CATTGCCCGA TACTGGCCAC CGCCAGCAGC ACGAACACAG GCTGGGCGCC
ACCCAGCAAC AGGCCGACCA GCAGCGCCTC CGACTTGCAG GGGCAAGCAG
GCTGGCGGCA CCGAAGGCAG AAAGAAACAG GCCGAAGTAG ACCGAAAAGT
CGAACACAGG TGCCATCCGG CAAAAAGTCG GG

Fig. 3E

Alignment of *Pseudomonas alcaligenes* repeat (PAR) elements from Contig 1

```
                All PARs     .TAAC.A.....T.A.....C...C.C......C..T..G.....C...............G........TTA....AA.CGTT...            (SEQ ID NO:96)
Consensus of    90% (16/18)  .TAAC.A...G.TCAA--..CG.TC.CT.CG.TC.CT..G.....C...............G.C....CCTTA.C..AA.CGTTAG.            (SEQ ID NO:97)
                Majority     CTAACAATTGGTTCAA--GTCGCTCGCTTCGCTCGCT-CGG-GACCG---GCTAAAG--CCGGCC---CCTTAACC-AAACGTTAGG            (SEQ ID NO:98)
                                      10        20        30        40        50        60        70        80

                PAR1         ........GC.C....--AG.....A.....T.....G-..--....---.-G..-----.C.GGCC.....G.TT..T......A      73 (SEQ ID NO:99)
                PAR2         .....T.C........--....T..........A..-...-....---........--.......---.........-........C      74 (SEQ ID NO:100)
                PAR3a        ........GC.C....CT......A.....T.....G-.ACAG..AAAA....CGCTTTT.T.TGCC.G...G.TT..T.......      86 (SEQ ID NO:101)
                PAR4         .......C........--.C.A..A..........-...-....---CG.TCC.--.-....---.........-.......G..      73 (SEQ ID NO:102)
                PAR5         -.........A.....--....T..........A..G..-....---........--.......---.........-.........      74 (SEQ ID NO:103)
                PAR6         .....T.C........--...A...........T.-...-....---CG.TCC.--.-....---.........-..........      73 (SEQ ID NO:104)
                PAR7         -..........C....--.C........C.....A..CG.AC.T...TAA....CGCTT......GCC.....G..-.........      83 (SEQ ID NO:105)
                PAR8         .......A........--....T..........A..-...-....---........--.......---.........-.........      74 (SEQ ID NO:106)
                PAR9         .....T...CAG....---G..GA...AAAC.C....G..-C.GT.-----------TTT.--.GCCG....T.TC..G......A      69 (SEQ ID NO:107)
                PAR10        ................--.C...............-...-..T..---..-G...--.....A---.........-.........A      73 (SEQ ID NO:108)
                PAR11        ...............G--A..T..........A..G..-....---...G...--.......---.........-..........      75 (SEQ ID NO:109)
                PAR12        .....T.C........--....T..........A..-...-....---........--.......---.........-.........      74 (SEQ ID NO:110)
                PAR13        .......A........--.C..T..........A..-...-....---........--.......---.........-........A      74 (SEQ ID NO:111)
                PAR14        ........GC.C....--AG.....A.....T.....G-..--.T..---........-----.C.GGCC.....G.TT..T.....      73 (SEQ ID NO:112)
                PAR15a       ........GC.C.T..CT......A.....T.....G-.ATAGT.AAAA...GCGCTTTT.T.TGTC.G...G.TT..T.....      85 (SEQ ID NO:113)
                PAR16        .......C........--A.........C.......--..-.....---..-.T.--.......---.........-..G......A      72 (SEQ ID NO:114)
                PAR17        .............A..--AC..T..........A..--..-.....---........--.......---.........-.........      73 (SEQ ID NO:115)
                PAR18        -......C........--.C..............A..-...-....---........T--T....---.....GG.-........A      72 (SEQ ID NO:116)
```

DOTPLOT of: superintegronI.pnt Density: 16075.00 May 25, 1999 16:22

COMPARE Window: 21 Stringency: 14 Points: 52,932

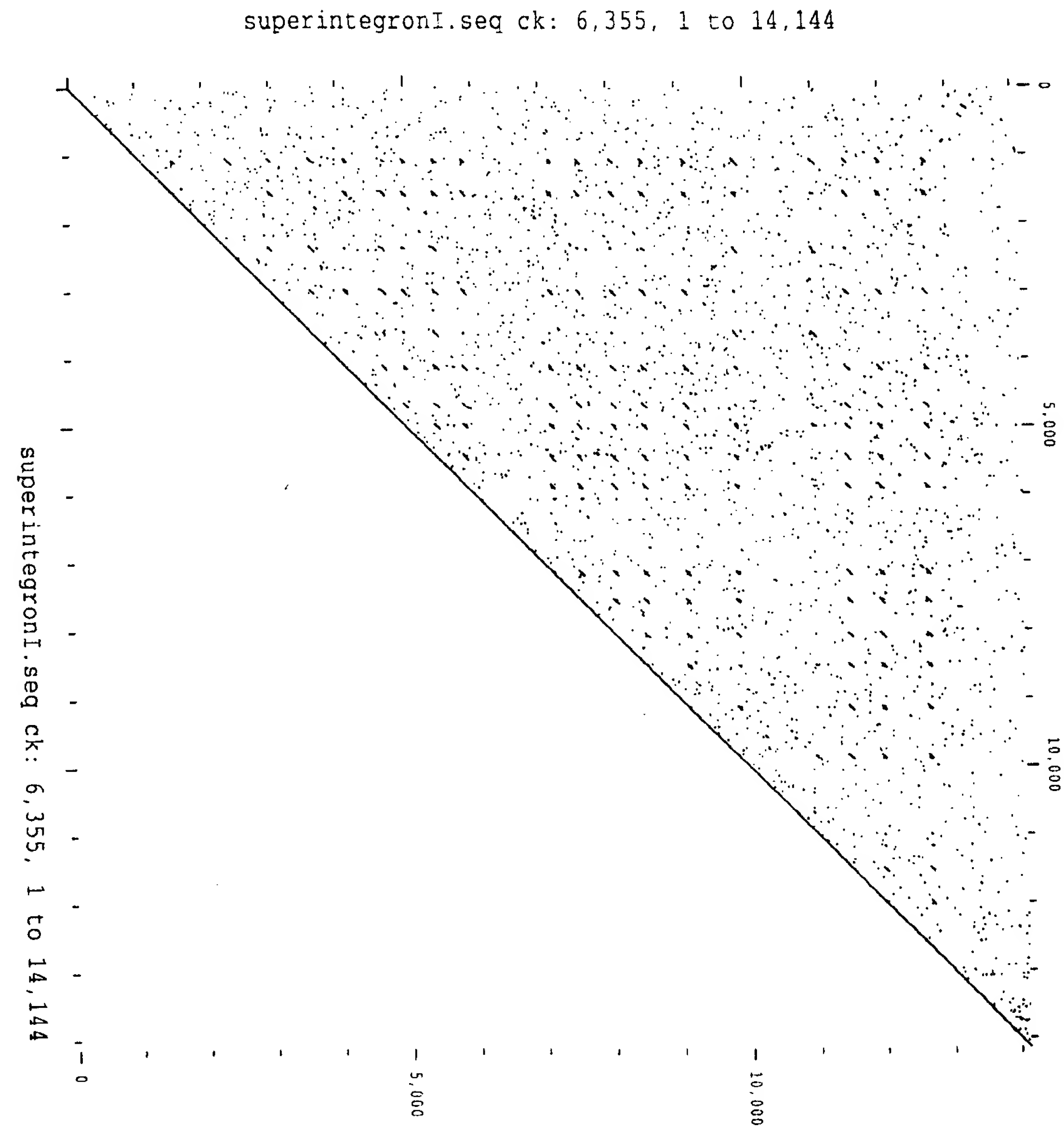


Fig 7

Fig 5

Squiggle plot of: Parla.mfold May 26, 1999 11:38

(Linear) MFOLD of: Parla T: 37.0 Check: 3607 from: 1 to: 78 May 26, 1999 11:38

Length: 78 Energy: -16.4

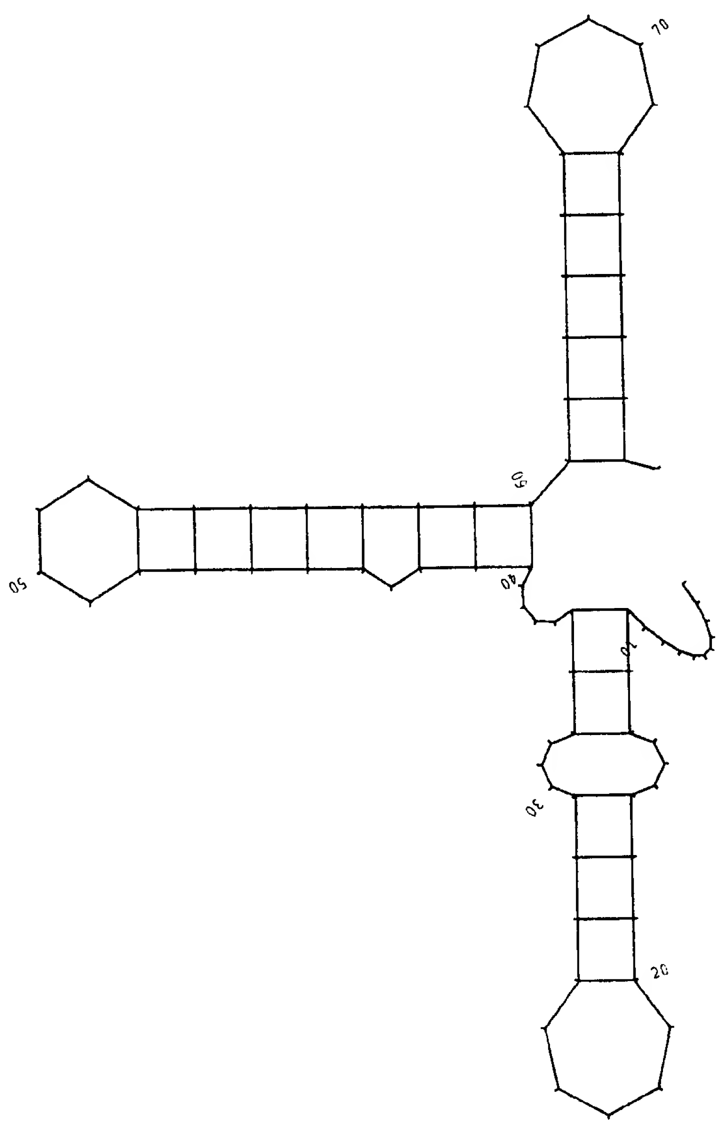


Fig. 6A

Family 1 of *Pseudomonas alcaligenes* repeat (PAR) elements

```
Identity      ...TAAC.......T.A..........CT.C..TC.CT........C..G...A.GCC.....GC.CC..T...C....CGTTA...   (SEQ ID NO:117)
90% (24/27(   ..CTAAC.A.TGGTTCAA........GCTTCGCTCACT.--.GGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGG.   (SEQ ID NO:118)
Majority      GCCTAACAATTGGTTCAA-GCCGTTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:119)

PAR1          ATCTAACAATTGGTTCAA-GTCGCTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAGCGTTAGGT   (SEQ ID NO:5)
PAR2          ACCTAACATGGCGCTCAACCGCGCTCCCTTCGGTCGCTGGACGCTGCGCGATAAAGCCGCGCAGCGCCGGTTAGCTCTACGTTAGGC   (SEQ ID NO:6)
PAR7          ACCTAACAACTGGTTCAA-GTCGTTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:11)
PAR8          GCCTAACAATTGGCTCAAG-TCGTTCGCTTCGCTCACT---CGGGACCGGC-GAAGCCG----GCCCCT-TAG-CCAAACGTTAGGT   (SEQ ID NO:12)
PAR10         GCCTAACAAGTGGTTCAAA-CCGTTCGCTTCGCTCACTG----GGACGGGCTAAAGCCC----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:14)
PAR14         CCCTAACAAATGGTTCAAAGCCGTTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGAG   (SEQ ID NO:18)
PAR15         CTCTAACAAATGGTTCAA-GTCGCTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:19)
PAR18         CCCTAACTAGTGGTTCAA-GCCGCTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:22)
PAR19         GCCTAACAAATGGTTCAAGTCGTTC-GCTTCGCTCACTG--CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGT   (SEQ ID NO:23)
PAR20         ACCTAACAATTGGTTCAA-GTCGTTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:24)
PAR22         GCCTAACAATGCGTTCAA-GTCGTTCGCTTCACTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGT   (SEQ ID NO:26)
PAR34         TTATAACAATTGGTTCAAGTCGTTC-GCTTCGCTCACTG--CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:38)
PAR36         GCCTAACAATAGGTTCAA-GTCGCTCGCTTCGCTCACT---TGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGT   (SEQ ID NO:40)
PAR37         ACCTAACTAGTGGTTCAA-GCCGCTCGCTTCGCTCACT---CGGGACCGGCTAAAGCCG----GCCCCT-TAA-CCAAACGTTAGGC   (SEQ ID NO:41)
PAR38         ATATAACAATTGGTTCAAA-CCGTTCGCTGCGCTCACTG----GGACGGGCTAAAGCCC----GCCCCT-TAA-CCAAACGTTATGC   (SEQ ID NO:42)
```

Fig. 6B
Family 2 of *Pseudomonas alcaligenes* repeat
(PAR) elements

Identity	....AAC.....G.TCAA...C.CTC.C...G.TC..T.G.....G..C........CG....GC.C...TTA......CGTT....	(SEQ ID NO:120)
90% (13/14)	....AACA...GGTTCAA...C.CTC.CT.CGCTC..T.GGGACCG-.C--......CG----GC.CC--TTA.CCAAACGTT.G..	(SEQ ID NO:121)
Majority	GCCTAACAACTGGTTCAAG-TCGCTCGCTTCGCTCGCTCGGGACCG-GC--TAAAGCCG----GCCCC--TTAACCAAACGTTAGGG	(SEQ ID NO:122)
PAR2	ACCTAACATGGCGCTCAACCGCGCTCCCTTCGGTCGCTGGACGCTGCGCGATAAAGCCGCGCAGCGCCGGTTAGCTCTACGTTAGGC	(SEQ ID NO:6)
PAR17	GCCCAACAAATGGTTCAAG-TCGCTCGCTCCGCTCACTCGGGACCG-GC--TAAAGCCG----GCCCC--TTAACCAAACGTTAGGG	(SEQ ID NO:21)
PAR21	GCCTAACAACTGGTTCAAG-TCACTCGCTTCGCTCGTTCGGGACCG-GC--ATAGC-CG----GCCCC--TTAACCAAACGTTAGGT	(SEQ ID NO:25)
PAR28	GCCTAACAACTGGTTCAAG-CCACTCGCTTCGCTCGCTCGGGACCG--C--GTACCGCG----GCCCC--TTAACCAAACGTTGGGC	(SEQ ID NO:32)
PAR29	GCCCAACAAACGGTTCAAGACCGCTCGCCTTGCTCGCTCGGGACCG-GC--TAAAACCG----GCCCC--TTAACCAAACGTTAGGG	(SEQ ID NO:33)
PAR30	GCCTAACAACTGGTTCAA-ATCGCTCGCTCCGCTCGCT-GGGACCG-GC---GAAGCCG----GCCCC--TTAACCAAACGTTAGGC	(SEQ ID NO:34)
PAR44	CCCTAACAAATGGTTCAAG-TCACTCGCTTCGCTCGTTCGGGACCG-GC--TAAAGCCG----GCCCC--TTAACCAAACGTTAGAG	(SEQ ID NO:48)
PAR53	GCCTAACAACTGGTTCAAG-CCACTCACTTCGCTCGCTCGGGACCG--C--GTTCCGCG----GCCCC--TTAACCAAACGTTGGGC	(SEQ ID NO:57)
PAR56	GCCTAACTACTGGTTCAAG-TCACTCGCTTCGCTCGTTCGGGACCG--C--GTTCCGCG----GCCCC--TTAACCAAACGTTAGGC	(SEQ ID NO:60)
PAR61	ATCTAACATGTGGTTCAAG-CCGCTCGCTTCGCTCACTCGGGACCG-GC--TAAAGCCG----GCCCC--TTAACCAAACGTTAGAG	(SEQ ID NO:65)
PAR62	CTCTAACAATTGGTTCAAG-CCGCTCGCTTCGCTCGCTCGGGATCG-GC---GAAGCCG----GCACC--TTAACCAAACGTTAGAG	(SEQ ID NO:66)
PAR67	GCCTAACAACTGGTTCAA-ATCGCTCGCTCCGCTCGCT-GGGACCG-GC---ATAGCCG----GCCC---TTAACCAAGCGTTAGAT	(SEQ ID NO:71)
PAR69	GTTTAACAACTGGTTCAAG-CCGCTCGCTTCGCTCACTCGGGACCG-GC--TAAATTCG----GCCCC--TTAGGCAAACGTTAACT	(SEQ ID NO:73)
PAR71	GTATAACAATTGGTTCAAG-TCACTCGCTTCGCTCGCTCGGGACCG-GC--TAAAGCCG----GCCCC--TTAACCAAACGTTAGAT	(SEQ ID NO:75)

Fig. 6C
Family 3 of *Pseudomonas alcaligenes* repeat (PAR) elements

Identity	....AACA ..CGCT.AAC...CGCTC.CT.CG.TCGCTGGA..............GC........G.C.GTTAGCT..A.CGTTA...	(SEQ ID NO:1[illegible])
90% (13/15)	..CTAACAATGCGCTCAACT..CGCTCACTTCGTTCGCTGGAC...CAAAAGCT.CGCTTTTG...GCCCGTTAGCTTAATCGTTA...	(SEQ ID NO:1[illegible])
Majority	ACCTAACAATGCGCTCAACTGTCGCTCACTTCGTTCGCTGGACAGCCAAAAGCTGCGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:125)
	10 20 30 40 50 60 70 80	
PAR5	ACCTAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACAGTCAAAAGCTGCGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:9)
PAR9	ACCTAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACAGTCAAAAGCTGCGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:13)
PAR32	ACCTAACAATGCGCTCAACTGTCGCTCACTTCGTTCGCTGGACAGTCAAAAGCTGCGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGCG	(SEQ ID NO:36)
PAR41	GCCTAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACAGTCAAAAGCTGCGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGAG	(SEQ ID NO:45)
PAR25	ACATAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACAGCCAAAAGCTACGCTTTTGCCTGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:29)
PAR13	ATTTAACAATGCGCTCAACTGTCGCTCACTTCGTTCGCTGGACAGCCAAAAGCTGCGCTTTTGTCTGCCCGTTAGCTTAATCGTTAGGG	(SEQ ID NO:17)
PAR52	GGCTAACAATGCGCTCAACTGTCGCTCACTTCGTTCGCTGGACAGCCAAAAGCTACGCTTTTGTCTGCCCGTTAGCTTAATCGTTAGGA	(SEQ ID NO:56)
PAR66	TGCTAACAATGCGCTTAACTGTCGCTCACTTCGTTCGCTGGATAGTCAAAAGCTGCGCTTTTGTCTGTCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:70)
PAR26	GCCTAACAAGTCGCTCAACTGCCGCTCACTCCGTTCGCTGGACAGCCAAAAGCTGCGCTTTTGTCTGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:30)
PAR24	GCCTAACAATACGCTCAACTATCGCTCACTTCGTTCGCTGGACGTCCAAAAGCTGCGCTTTTGGCCGCCCGTTAGCTTAACCGTTATGT	(SEQ ID NO:28)
PAR27	GCCTAACAATGCGCTCAACTATCGCTCACTCCGTTCGCTGGACGTCCAAAAGCTGCGCTTTTGGCCGCCCGTTAGCTTAATCGTTAGGC	(SEQ ID NO:31)
PAR54	GCCCAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACGTCCAAAAGCTACGCTTTTGGCCGCCCGTTAGCTTAATCGTTATGC	(SEQ ID NO:58)
PAR4	ACCTAACAATGCGCTCAACTGCCGCTCACTTCGTTCGCTGGACTCGCAAAAGCTGCGCTTTTGCTCGCCCGTTAGCTTAATCGTTAGGT	(SEQ ID NO:8)
PAR11	CTCTAACAATGCGCTCAACTATCGCTCACTTCGTTCGCTGGACTCGCAAAAGCTGCGCTTTTGCTCGCCCGTTAGCTTAATCGTTATAA	(SEQ ID NO:15)
PAR2	ACCTAACATGGCGCTCAACCG-CGCTCCCTTCGGTCGCTGGACGCTGCGCGATAAAGCCGCGCAGCGCCGGTTAGCTCTA-CGTTAGGC	(SEQ ID NO:6)

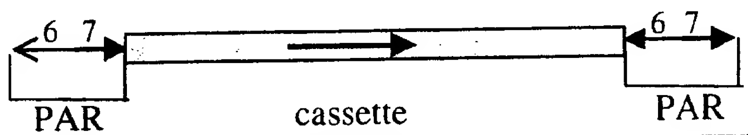
Fig. 6D
Family 4 of *Pseudomonas alcaligenes* repeat (PAR) elements

```
Identity   ...TAACA...CGCTCAA...CG.TC.CT.CG.T.C.CTGG.....G.GC....AAGCC.....GC.CC..TTAGCT....CGTTA...  (SEQ ID NO:126)
90% (7/8)  ...TAACAA.GCGCTCAAA-.CGCTC.CTTCG.T-CGCTGGG-AC.G-GC--..AAGCC.----GCCCC--TTAGCTTAATCGTTAG..  (SEQ ID NO:127)
Majority   GCCTAACAATGCGCTCAAA-GCGCTCACTTCGTT-CGCTGGG-ACCG-GC--TAAAGCCG----GCCCC--TTAGCTTAATCGTTAGGT  (SEQ ID NO:128)
                    10        20        30        40        50        60        70        80

PAR6       GCCTAACAATGCGCTCAAA-GCGCTCACTTCGTT-CGCTGGG-ACCG-GC---GAAGCCG----GCCCC--TTAGCTTAATCGTTAGGT  (SEQ ID NO:10)
PAR31      GCCTAACAATGCGCTCAAA-GCGCTCACTTCGTT-CGCTGGG-ACCG-GC---GAAGCCG----GCCCC--TTAGCTTAATCGTTAGGT  (SEQ ID NO:35)
PAR50      GCCTAACAATGCGCTCAAA-GCGCTCACTTCGTT-CGCTGGG-ACCG-GC---GAAGCCG----GCCCC--TTAGCTTAATCGTTAGAA  (SEQ ID NO:54)
PAR65      GTCTAACAATGCGCTCAAA-GCGCTCACTTCGTT-CGCTGGG-ATCG-GC--TAAAGCCG----GCCCC--TTAGCTTAATCGTTAGCA  (SEQ ID NO:69)
PAR42      CTCTAACAATGCGCTCAAA-TCGCTCACTACGTT-CGCTGGG-ACCG-GC--TAAAGCCG----GCCCC--TTAGCTTAATCGTTAGAG  (SEQ ID NO:46)
PAR72      GCATAACAAGTCGCTCAAA-TCGCTCACTTCGTTTCGCTGGG-ACGG-GC--TAAAGCCC----GCCCC--TTAGCTTA-TCGTTAGGC  (SEQ ID NO:76)
PAR12      TTATAACAATGCGCTCAAA-TCGTTCGCTTCGCT-CACTGGG-ACGG-GC--TAAAGCCC----GCCCC--TTAGCTTAATCGTTAAAT  (SEQ ID NO:16)
PAR2       ACCTAACATGGCGCTCAACCGCGCTCCCTTCGGT-CGCTGGACGCTGCGCGATAAAGCCGCGCAGCGCCGGTTAGCTCTA-CGTTAGGC  (SEQ ID NO:6)
```

Fig 7A. PAR-specific oligonucleotide (bottom) aligned with PAR majority consensus (top)

```
majority   GCCTAACAATTGGTTCAAG-GTCGCTCGCTTCGCTCACT-CGGGACCGGCTAAAGC--------CGGCC--CC-TTAA-CCAAACGTTAGGC (SEQ ID NO:129)

oligo 1                   5' TCGCTTCGCTCACTGCGGGACCGGCTAAAGC--------CGGCC--CC-TTAA-CCAAACGTTA     (SEQ ID NO:79)
      2    5' TAACAATTGGTTCAAG--TCGTTCGCTTCGCTCACTGCGGGACCG                                    (SEQ ID NO:80)
      3    5' TAACTATTCAGTCAAGCGGA---CGCAAACCCCGCTGCGCGGTCTT                                   (SEQ ID NO:81)
      4    5' TAACAATGCGCTCAACTGCG-CTCACTTCGTTCGCTGGACAGCC                                     (SEQ ID NO:82
      5    5' TAACAAGTCGCTCAACTGCCGCTCACT-CGTTCGCTGGACAGCC                                     (SEQ ID NO:83)
```

Fig. 7B. PCR primers for PAR fingerprints




```
PAR majority:                 5' GCCTAACAATTGGTTCAAG-GTCGCTCGCTTCGCTCACT-CGGGACCGGCTAAAGC   (SEQ ID NO:129)
Oligo 6 (complement of PAR):                      3' GCGAGCGAAGCGAGTGA-GCC                  (SEQ ID NO:84

PAR majority (cont'd):                            --------CGGCC--CC-TTAA-CCAAACGTTAGGC      Continuation of
                                                                                            (SEQ ID NO:129)
Oligo 7:                                                  GCC--CC-TTAA-CCAAACGTTAG 3'       (SEQ ID NO:85)
```

Fig 8. Hybridization of a PAR-specific Oligonucleotide 1 to *Pseudomonas alcaligenes* chromosomal DNA.

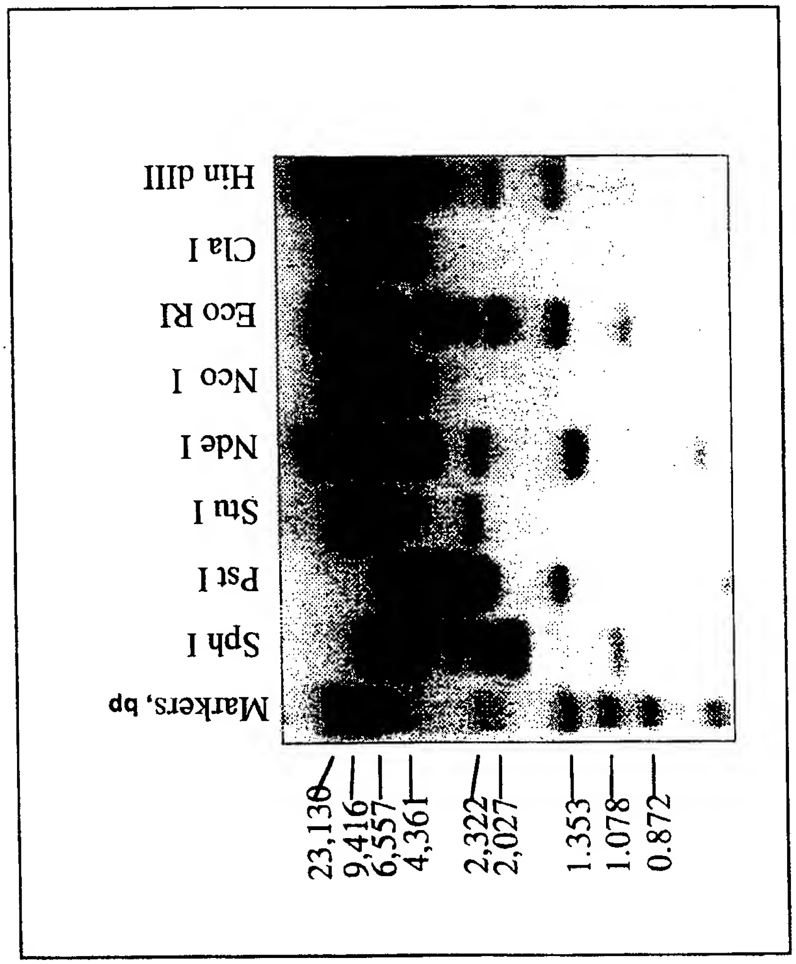
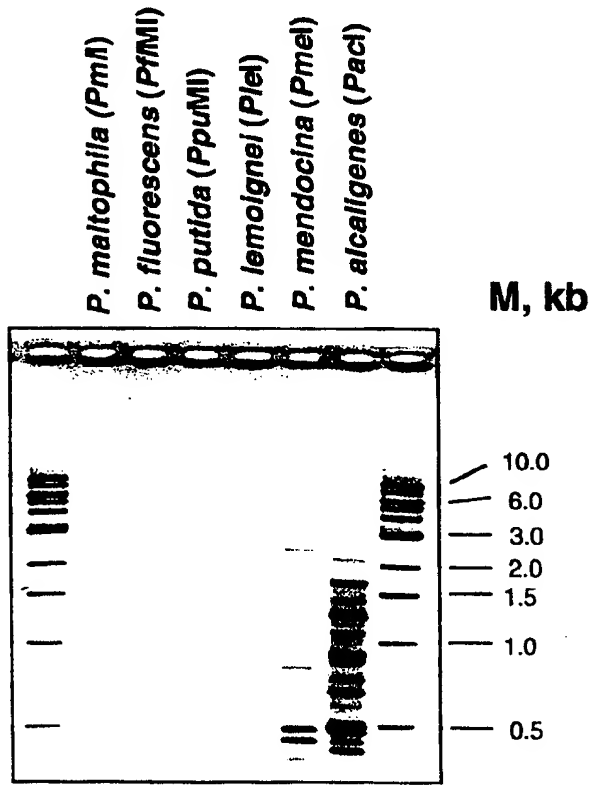


Fig.9. Distribution of PAR Cassettes Among *Pseudomonas* Species



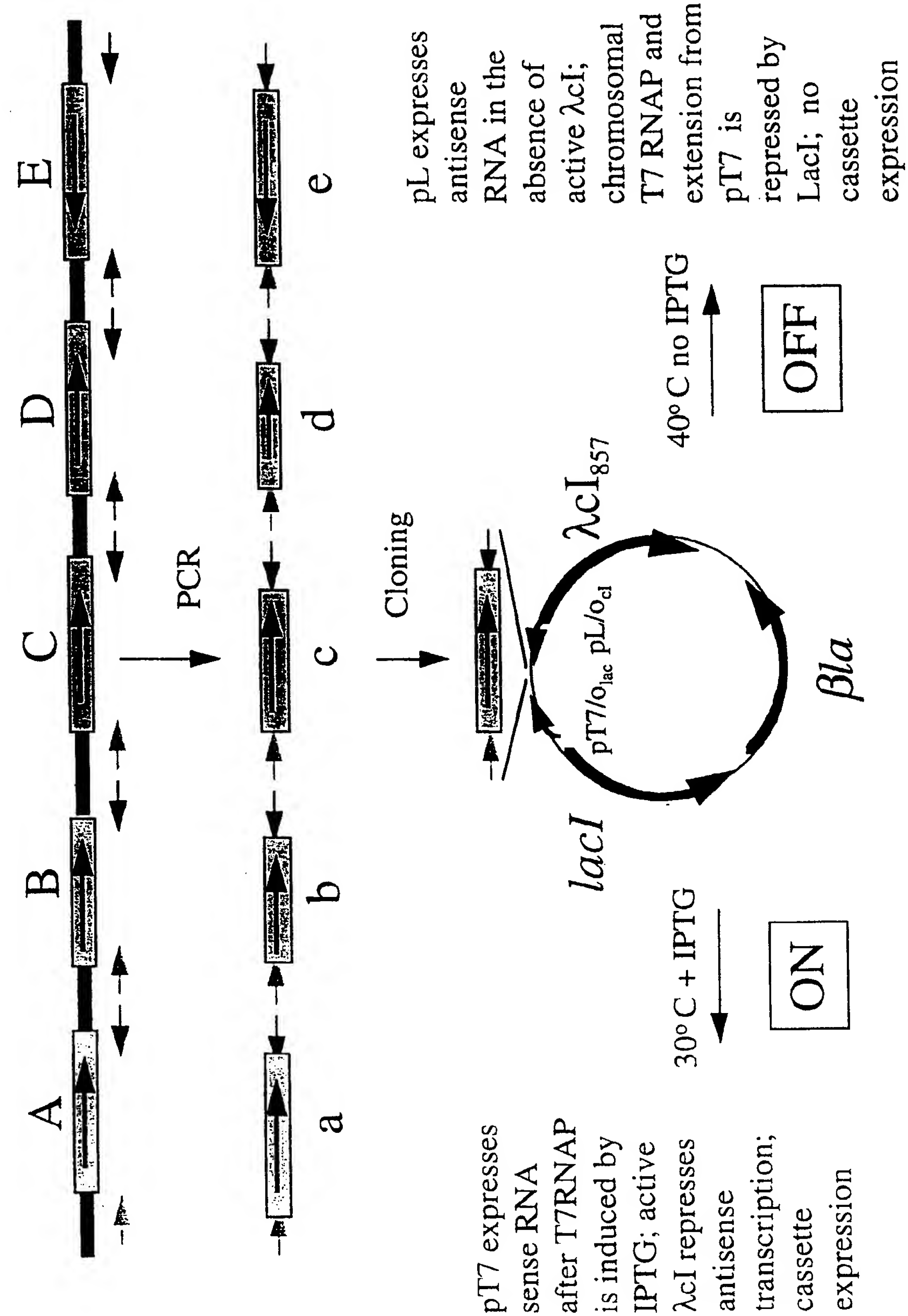


Fig. 10. Amplification and cloning strategy